EPA/600/R-01/071
September, 2001

Probabilistic Aquatic Exposure Assessment for Pesticides

I: Foundations

By

Lawrence A. Burns, Ph.D.
Ecologist, Ecosystems Research Division
U.S. Environmental Protection Agency
960 College Station Road
Athens, Georgia 30605-2700

National Exposure Research Laboratory
Office of Research and Development
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

Notice

The U.S. Environmental Protection Agency through its Office of Research and Development funded and managed the research described here under GPRA Goal 4, *Preventing Pollution and Reducing Risk in Communities, Homes, Workplaces and Ecosystems*, Objective 4.3, *Safe Handling and Use of Commercial Chemicals and Microorganisms*, Subobjective 4.3.4, *Human Health and Ecosystems*, Task 6519, *Advanced Pesticide Risk Assessment Technology*. It has been subjected to the Agency's peer and administrative review and approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Abstract

Models that capture underlying mechanisms and processes are necessary for reliable extrapolation of laboratory chemical data to field conditions. For validation, these models require a major revision of the conventional model testing paradigm to better recognize the conflict between model user's and model developer's risk (as Type I and Type II errors) in statistical testing of model predictions. The predictive reliability of the models must be hypothesized and tested by methods that lead to conclusions of the form "the model predictions are within a factor-of-two of reality at least 95% of the time." Once predictive reliability is established, it can be treated as a "method error" within a probabilistic risk assessment framework. This report, developed under APM 131 ("Develop a Probability-Based Methodology for Conducting Regional Aquatic Ecosystem Exposure and Vulnerability Assessments for Pesticides"), describes a step-by-step process for establishing the predictive reliability of exposure models.

Monte Carlo simulation is the preferred method for capturing variability in environmental driving forces and uncertainty in chemical measurements. Latin Hypercube Sampling (LHS) software is under development to promote efficient computer simulation studies and production of tabular and graphical outputs. Desirable outputs include exposure metrics tailored to available toxicological data expressed as distribution functions (pdf, cdf) and, if needed, empirical distribution functions suitable for use in Monte Carlo risk assessments combining exposure and effects distributions. ORD numerical models for pesticide exposure supported under this research program include a model of spray drift (AgDisp), a cropland pesticide persistence model (PRZM), a surface water exposure model (EXAMS), and a model of fish bioaccumulation (BASS). A unified climatological database for these models is being assembled by combining two National Weather Service (NWS) products: the Solar and Meteorological Surface Observation Network (SAMSON) data for 1961-1990, and the Hourly United States Weather Observations (HUSWO) data for 1990-1995. Together these NWS products provide coordinated access to solar radiation, sky cover, temperature, relative humidity, station atmospheric pressure, wind direction and speed, and precipitation. By using observational data for the models, "trace-matching" Monte Carlo simulation studies can transmit the effects of environmental variability directly to exposure metrics, by-passing issues of correlation (covariance) among external driving forces. Additional datasets in preparation include soils and land-use (planted crops) data summarized for the State divisions of Major Land Resource Areas (MLRA), derived from National Resource Inventory (NRI) studies.

This report covers a period from May 2, 2001 to September 30, 2001 and work was completed as of September 30, 2001.

Preface

Predictive modeling is an important tool for assessing the environmental safety of new pesticidal active ingredients and new uses for currently registered products, and for evaluating the implications of new findings in their environmental and product chemistries. Climate, soil properties, limnology, and agronomic practices influence exposure by controlling the movement of pesticides within the agricultural landscape and by governing the speed and products of transformation reactions. These factors vary with time and with location within the often continent-wide use patterns of agricultural chemicals. This variability, together with measurement uncertainties in the values of chemical properties, mandates a statistical and probabilistic approach to exposure assessment. An effective pesticide modeling technology must include validated algorithms for transport and transformation of pesticides, extensive databases of agro-ecosystem scenarios (crop and soil properties, meteorology, limnology, fish community ecology) and graphical user interfaces to maximize the ease of production and interpretation of complex probabilistic analyses. Several agencies collect data of significance for environmental safety, but these data must be assembled in usable forms, organized by appropriate landscape units, and made accessible to simulation models if their potential is to be realized.

This report is a foundational document for predictive modeling in support of pesticide exposure studies. The Environmental Protection Agency's approach to pesticide regulation in its Office of Pesticide Programs (OPP) is outlined, with a brief discussion of the development of probabilistic exposure assessment within OPP's Environmental Fate and Effects Division (EFED). The epistemological basis for predictive modeling in support of EFED modeling is reviewed, with a brief description of Monte Carlo methods for incorporating variability and uncertainty into process-based models. Within the Monte Carlo context, the predictive uncertainty of numerical models must be explicitly quantified; a method of model validation and quantification of predictive uncertainty developed for this project is reported here along with a brief review of prior validation studies of the models. Finally, source materials for databases under development to support "trace matched" input parameters for exposure modeling are briefly documented.

Lawrence A. Burns, Ecologist
Ecosystems Research branch

Foreword

Environmental protection efforts are increasingly directed toward preventing adverse health and ecological effects associated with specific chemical compounds of natural or human origin. As part of the Ecosystems Research Division's research on the occurrence, movement, transformation, impact, and control of environmental contaminants, the Ecosystems Assessment Branch studies complexes of environmental processes that control the transport, transformation, degradation, fate, and impact of pollutants or other materials in soil and water and develops models for assessing the risks associated with exposures to chemical contaminants.

Rosemarie C. Russo, Director
Ecosystems Research Division
Athens, Georgia

Contents

Abbreviations, Acronyms, and Symbols

α	Probability of Type I error; the probability of rejecting the null hypothesis H_0 when H_0 is in fact true; P {reject H_0|H_0}
β	Probability of type II error; the probability of accepting the null hypothesis H_0 when in fact the alternate hypothesis H_a is true; P {accept H_0|H_a}; probability of accepting H_0 when H_0 is in fact false
$(1-\beta)$	The probability of rejecting H_0 when it is in fact false; the *power function* of a test
$100 \times \alpha$	Significance level expressed as a percentage (e.g., 5%)
$100\ (1-\alpha)$	Confidence level expressed as a percentage (e.g., 95%)
AgDisp	*Ag*ricultural *Disp*ersion Model
BASS	*B*ioaccumulation and *A*quatic *S*ystem *S*imulator
d.f.	Degrees of freedom
ECOFRAM	*E*cological *C*ommittee *o*n *F*IFRA *R*isk *A*ssessment *M*ethods
EFED	*E*nvironmental *F*ate and *E*ffects *D*ivision of OPP
EPA	Environmental Protection Agency
EXAMS	*Ex*posure *A*nalysis *M*odeling *S*ystem
FGETS	*F*ood and *G*ill *E*xchange of *T*oxic *S*ubstances
FIFRA	*F*ederal *I*nsecticide, *F*ungicide, and *R*odenticide *A*ct (legislation governing pesticide registration in the United States)
μ	The true mean of a variable
ν	Degrees of freedom
NRC	National Research Council, National Academy of Sciences
OPP	Office of Pesticide Programs, U. S. Environmental Protection Agency
ORD	Office of Research and Development, U.S. Environmental Protection Agency
PRZM	*P*esticide *R*oot *Z*one *M*odel
σ^2	The variance of a variable
S^2,s^2	Sample estimate of variance
S, s, S.D.	Sample standard deviation
S.E.	Standard error of the mean
$\bar{x}$	Sample estimate of the mean

Acknowledgments

Development of the model validation technique described in this report benefitted from reviews by R.A. Ambrose, M.C. Barber, L.A. Suárez, and J. Babendreier. The descriptions of the current validation status of the AgDisp and BASS models were supplied by S.L. Bird and M.C. Barber respectively. L.A. Suárez prepared **Figure 1** and **Figure 2**; L. Prieto prepared **Figure 12** and **Figure 13**. Discussions with the OPP/EFED implementation team for probabilistic risk assessment (Kathryn Gallagher, James Lin, Leslie Touart, Ron Parker, et al.) were helpful in formulating the objectives of the work herein reported. **Figure 1**, **Figure 2** and **Figure 3** are from *Mathematical Modeling of Biological Systems* by Harvey J. Gold, copyright © 1977 by John Wiley & Sons, Inc. This material is used by permission of John Wiley & Sons, Inc.

Introduction to Pesticide Safety Assessment

Exposure assessment modeling is an important component of U.S. Environmental Protection Agency (EPA) ecological risk assessments for pesticides. Estimates of uncertainty in model results, although long recognized as desirable by the Office of Pesticide Programs (OPP) [1], have primarily been accommodated in the regulatory process by imposing safety factors and various conservative assumptions on modeling scenarios and exposure metrics. The Agency uses a "tiered" approach to risk assessment as a means of focusing attention on the most problematic pesticides and use patterns. This approach is one of screening out, via the use of conservative assumptions, pesticides posing minimal risk to non-target biota. Those materials failing to pass simple screening tests are remanded to higher tier, more complex risk analyses. Thus, the most conservative assumptions are not used to restrict usage or "ban" chemicals, but only to segregate materials according to the intensity of scrutiny they are to receive during the analysis phase of a risk assessment.

Recommendations and analyses of the National Research Council (NRC) have been formative in the development of OPP regulatory analyses and procedures. NRC's analysis of risk assessment in regulatory agencies [2] was adopted early by EPA as its standard approach to chemical safety issues. The NRC first separated *risk assessment* from *risk management*, in order to encourage clarification of the boundaries between the technical and scientific components of regulatory activities, as against the social, economic, and political pressures that constrain regulatory decision-making.

Risk assessment is the characterization of the potential adverse effects of exposure to environmental hazards. Risk assessments include several elements: description of the potential effects on organisms based on toxicological, epidemiological, and ecological research; extrapolation from those results to predict the type and estimate the severity and extent of effects in natural populations under given conditions of exposure; estimation of the species and locations of organisms exposed at various intensities and durations; and summary judgments on the existence and magnitude of ecological problems. The process includes both quantitative risk assessment, with its reliance on numerical results, and qualitative expressions or judgments made during model parameterization and in the evaluation of the uncertainties inherent in generalized (e.g., regional) exposure analyses and laboratory-to-field extrapolation of toxicological studies.

Risk management is the process of evaluating alternative regulatory actions and selecting among them. Risk management, which is carried out by EPA under its various regulatory mandates, is the Agency decision-making process that entails consideration of political, social, economic, and engineering information to develop, analyze, and compare possible regulatory responses to a potential ecotoxicological hazard. The selection process necessarily requires the use of value judgments on such issues as the acceptability of risk and the reasonableness of the costs of controls.

Steps in Risk Assessment

The NRC report divided risk assessment into four major steps: hazard identification, dose-response assessment, exposure assessment, and risk characterization. A risk assessment may stop with the first step, hazard identification, if no adverse effects are found, or if the Agency elects to take regulatory action without further analysis for reasons of policy or statutory mandate.

Hazard identification is the process of determining whether exposure to an agent can cause an increase in the incidence of a biological effect (mortality, reproductive impairment, etc.). It involves characterizing the nature and strength of the evidence of causation. In the case of pesticides, biological effects must be evaluated in the context both of efficacy (biocidal impact on the target organism), and the incidental endangerment of non-target organisms, both on-site and off-site following transport of the pesticide out of the intended target area. In the case of industrial chemicals and waste disposal operations, biological effects are universally undesirable. There are, however, few cases in which direct biological data are available. The question, therefore, is often restated in terms of the effects on laboratory animals or other test systems (microcosms, mesocosms), e.g., "Does the agent produce mortality in test animals?" Positive answers to such questions are typically taken as evidence that an agent may pose a risk to exposed natural systems. Information from short-term *in vitro* (e.g., bacteriological bioassay) tests and inferences from structural similarities to known chemical hazards may also be considered.

Dose-response assessment is the process of (1) characterizing the relation between the dose of an agent administered or received and the incidence of an adverse effect in an experimentally exposed population, and (2) of estimating the incidence of the effect in natural populations and ecosystems as a function of exposure to the agent. It takes account of intensity and duration of exposure, age patterns of exposure, and possibly other variables that might affect response such as sex, seasonal variation in condition, and other modifying factors. A dose-response assessment usually requires extrapolation from high to low dose and extrapolation from test species to potential target species. A dose-response assessment should describe and justify the methods of extrapolation used to predict incidence, and should characterize the statistical and biological uncertainties in these methods.

Exposure assessment, of special interest in the present context, is the process of measuring or estimating the intensity, frequency, and duration of contact of the biota with an agent currently present in the environment, or of estimating hypothetical exposures that might arise from the release of new chemicals into the environment. In its most complete form, it describes the magnitude, duration, schedule, and route of exposure; the size, condition, and species of the biological entities exposed; and the uncertainties in all estimates. Exposure assessment is often used to identify feasible prospective regulatory control options and to predict the effects of available control technologies on exposure. Quantitative exposure prediction is often required in EPA regulatory analyses. Goals may include an estimate of the effects of new chemicals prior to manufacture or from increased production volumes of existing chemicals, an evaluation of off-site impacts of pesticides during the initial registration process or prior to expansion of use areas, or establishment of remediation priorities among hazardous waste sites.

Risk characterization is the process of estimating the incidence of an effect under the various conditions of contamination of natural systems and biological contact described in the exposure assessment. It is performed by combining the exposure and dose-response assessments. The ultimate effects of the uncertainties in the preceding steps are described in this step. Within OPP, at lower tiers this step is usually conducted using a *quotient method*, i.e., a direct comparison of exposure and dose-response results using a ratio of the two to infer the magnitude of risk.

The OPP "Quotient Method" of Assessing Ecological Risk[1]

The EPA's Office of Pesticide Programs (OPP) adapted the NRC 1983 [2] paradigm to better fit its legislated mandates, goals, organizational structures and procedures. The goal of OPP ecological risk assessments is to provide the scientific basis needed to support and inform Agency risk management decisions and regulatory actions. These can range from registration of a pesticide, to placing restrictions on the permitted usages of a chemical, to requiring detailed laboratory or field testing of a chemical to better evaluate impacts, to outright prohibition of a chemical. Combining hazard identification and dose-response assessment into a single domain, OPP parsed chemical safety as a matter of *ecotoxicological hazard assessment*, *environmental exposure assessment*, and *ecological risk characterization.*

Ecotoxicological Hazard Assessment, in OPP practice, combines hazard identification and dose-response assessment into five steps leading to the ultimate goal of the assessment, a *toxicological level of concern*: the concentrations that, if equaled or exceeded in the environment, could reasonably be expected to produce adverse effects in the biota. This step-wise methodology encompasses five specific procedures:

1. Define the endpoints of concern. The term *endpoint* is used to denote the specific effects of pesticides that merit evaluation during risk assessment. Organisms exposed to a chemical may die, may fail to develop normally, or may fail to reproduce. Organisms may bioconcentrate chemicals to levels in their tissues that harm their predators[2] or the detrital food chain, irrespective of effects upon themselves. In addition, an early identification of the species potentially at risk can guide the assessment into the most effective avenues of investigation.

2. Select an appropriate test species for laboratory toxicological investigation. The myriad of potentially exposed species and ecosystems forces the use of experimental models, and the use of inferential rules to extrapolate from the laboratory to field exposure conditions. OPP has identified *surrogate species* for several categories of birds, mammals, fishes, reptiles, amphibians, invertebrates, and plants. The ideal test species is sensitive to chemical insult, ecologically ubiquitous with a well-known life history that accommodates the constraints of toxicological investigations, is highly valued by society, and is easily and inexpensively cultured in toxicological laboratories. The objective is to avoid the massive uncertainties and controversy that plague human health assessments that must extrapolate from mouse to man: if the test organism is *per se* of concern, then the laboratory toxicology is directly relevant. The surrogates chosen by OPP – rainbow trout, mallard duck, etc.– fit these criteria rather well. Still, in many instances toxicology must be inferred. For example, it is not generally feasible to test endangered species directly for toxicological impact, although some endangered species have been cultured by the US Fish and Wildlife Service for just such purposes (Foster L. Mayer, Jr., personal communication).

[1] Developed from Office of Pesticides and Toxic Substances (1990). State of the Practice Ecological Risk Assessment. US Environmental Protection Agency, Washington, DC.

[2] One obvious example is the closing of commercial fisheries due to contamination of the fishes with PCBs.

3. OPP uses a *tiered testing system* of four progressively more complex and expensive tests. Testing begins with short-term acute and sub-chronic laboratory studies; those most commonly conducted develop basic dose-response data (the LC_{50} and the LD_{50}). The second and third tiers usually include an expanded series of acute and chronic tests for a wider variety of organisms, as well as fish bioaccumulation factors. The fourth tier may include field tests and mesocosm or pond studies.

4. Validation of test data is so important to sound environmental regulation that OPP regards it as a separate sub-discipline of ecotoxicological hazard assessment. The National Research Council [2] listed among the aims of dose-response assessment the characterization of "statistical and biological uncertainties." The OPP analysis of test data evaluates *accuracy* (e.g., appropriate test conditions, interfering collateral contaminants), *precision* (adherence to prescribed methods yields repeatability within about a factor of two), and *sensitivity* (the design and conduct of a test must be adequate to detect the endpoint, i.e., the probability of a false negative must be evaluated). The Office of Pesticide Programs (OPP) has developed, and made publicly available, Standard Evaluation Procedures (*SEP*) for almost every kind of data required for OPP hazard assessments.

5. Finally, from the foregoing steps, OPP analysts develop a *toxicological level of concern*. This is the concentration that is compared to environmental exposures to produce a risk quotient; it is the concentration that reasonably can be expected to cause adverse effects. Although ideally the measured LC_{50} can be used directly, it often must be extrapolated from test subjects to additional species in order to provide a *margin of exposure* or safety margin (OPP); or to generate an Office of Pollution Prevention and Toxics (OPPT) *assessment factor*.

Environmental Exposure Assessment develops information of two kinds, essentially in conformance with the NRC [2] exposition:

! The intensity, duration, and frequency of contact between the biota and a potentially harmful agent – variously known as Expected Environmental Concentrations (EEC), Estimated Environmental Concentrations (EEC), Predicted Environmental Exposure or Predicted Environmental Concentration (PEC), etc.

! A profile (size, condition, species) of organisms potentially exposed to a chemical, and their distribution in the environment.

1. Estimating environmental concentrations requires several procedures. For pesticides, registrants submit information detailing target crops, application rates, frequency, timing, method, etc. The Pesticide Root Zone Model (PRZM) [3] calculates the transport off-site of pesticides and transformation products from the specifics of climate, soils, and crops. A direct interface between PRZM and EXAMS surface water models [4] collects the PRZM edge-of-field pesticide export data and converts them into EXAMS Mode 3 input load sequences.

OPP also constructs more direct estimates of chemical loadings on aquatic systems: some herbicides are applied directly to water bodies, and OPP frequently estimates the mass of pesticide entering water bodies due to spray drift. Exposure estimates of several kinds are performed in support of the risk assessment tiers (see Text Box 1, adapted from [5]).

Ecological Risk Characterization is the final component of ecological risk assessment in OPP chemical safety analyses. Risk characterization compares the toxicological levels of concern

Preliminary exposure analysis includes simple laboratory tests and models to provide an initial fate profile for a pesticide (hydrolysis and photolysis in soil and water, aerobic and anaerobic soil metabolism, and mobility).

Fate and transport assessment provides a comprehensive profile of the chemical (persistence, mobility, leachability, binding capacity, transformation products (metabolites and "degradates")) and may include field dissipation studies, published literature, other field monitoring data, ground-water studies, and modeled surface water estimates.

Estimated environmental concentrations (EEC) are derived during the exposure analysis or comprehensive fate and transport assessment. There are four EEC estimation procedures:

Level 1: A direct-application, high-exposure model designed to estimate direct exposure to a nonflowing, shallow-water (<15 cm) system.

Level 2: Adds simple drift or runoff exposure variables such as drainage basin size, surface area of receiving water, average depth, pesticide solubility, surface runoff, or spray drift loss, which attenuate the Level 1 direct application model estimate.

Level 3: Computer runoff and aquatic exposure simulation models. A loading model (SWRBB-WQ, PRZM, etc.) is used to estimate field losses of pesticide associated with surface runoff and erosion; the model then serves as input to a partitioning model (EXAMS) to estimate sorbed and dissolved residue concentrations. Simulations are based on either reference environment scenarios or environmental scenarios derived from typical pesticide use circumstances.

Level 4: Stochastic modeling where EECs are expressed as exceedence probabilities for the environment, field, and cropping conditions.

Text Box 1. Generalized exposure analysis methods and procedures used in prospective ecological risk screens of pesticides [5]

with the EECs developed in exposure assessment to judge whether there is sufficient risk to warrant further investigations or regulatory action. There are four steps in risk characterization:

First, the quotient of the EEC and the toxicological level of concern (TLC) is calculated to arrive at the *quotient* of the "quotient method": EEC / TLC = Quotient. Quotients>1 imply that frank effects are likely and regulatory action is indicated. Quotients<<1 (e.g., 0.01) imply risk is slight and little or no action is required. Quotients near 1 represent uncertainty in the risk estimate and usually require additional data. The second step of OPP risk characterization is to *compare the quotient to regulatory criteria*. Third, OPP may augment its analyses with *confirmatory data from mesocosm and field studies, incident report observations of mortalities, and ecological simulation models*. The final step in OPP ecological risk assessment is an evaluation of the *weight of the evidence*, that is, a review of the quotient method analyses and additional evidence available from field tests, incident reports, and simulation models. This step includes evaluation of the quality of all available data and the frequency and magnitude of effects in various media, while retaining the flexibility to include all available relevant scientific information in the final risk assessment.

Although predating their formulation, the ecological risk assessment methods used by OPP are consistent with EPA's [6] risk assessment guidelines. Two pesticide studies (carbofuran and synthetic pyrethroids) were influential during the development of the guidelines [7]. OPP has a continuing interest in the further development of probabilistic exposure analysis, methods for predicting the efficacy of risk mitigation measures, and the development of a set of standard data bases for consistent parameterization of fate and transport models [5, 8].

Probabilistic Assessment

OPP regards the quotient method as an entrenched, useful technology with numerous acknowledged weaknesses, including deficiencies in evaluation of indirect effects, disregard of incremental dose impacts, and neglect of effects at higher levels of organization. It nonetheless continues to provide a useful lower tier deterministic "screening method" when coupled to appropriately conservative assumptions. The exposure portion of these assessments has, however, been conducted for some time with explicit accounting for environmental variability within the limited scenarios employed. In these analyses, the PRZM/EXAMS models are run with between 22 and 36 years of input meteorological data, and the 90th percentiles of several exposure metrics are captured from their probabilistic plotting position. Although more informative than a simple dilution calculation, a fuller exposition of the properties of the distribution, with expansion of the analysis to include a variety of physiographic regions, soils, and agricultural landscapes, could serve to place the current point estimates in a fuller national, multi-use context. In a congressionally mandated review of EPA risk assessment procedures [9], the National Academy of Sciences strongly criticized EPA's approach to risk assessment of hazardous air pollutants insofar as "it does not supplant or supplement artificially precise single estimates of risk ('point estimates') with ranges of values or quantitative estimates of uncertainty...This obscures the uncertainties inherent in risk estimation, although the uncertainties themselves do not go away...without uncertainty analysis it can be quite difficult to determine the conservatism of an estimate." In 1997, the EPA established its *Policy for Use of Probabilistic Analysis in Risk Assessment* [10] and published a set of "guiding principles" for conducting such analyses using Monte Carlo methods [11]. The policy reaffirms the place of deterministic methods in the suite of Agency methods: "[Probabilistic] analysis should be a part of a tiered approach to risk assessment that progresses from simpler (e.g., deterministic) to more complex (e.g., probabilistic) analyses as the risk management situation requires." More importantly, the policy statement establishes a set of "conditions for acceptance" by the Agency of probabilistic analyses. These conditions, intended to encourage the ideals of transparency, reproducibility, and the use of sound methods, identify factors to be considered by Agency staff in implementing the policy (see page 5).

In May of 1996, the OPP EFED (Environmental Fate and Effects Division) presented two ecological risk assessment case studies to its FIFRA Scientific Advisory Panel (SAP) for comment. The SAP affirmed the value of the process, but urged that OPP begin development of tools and methods for probabilistic assessments. In response, EFED instituted an "Ecological Committee on FIFRA Risk Assessment Methods" (ECOFRAM) composed of four workgroups (aquatic and terrestrial exposure and effects). ECOFRAM was composed of experts drawn from government agencies, academia, environmental groups, industry, and other stakeholders. Following completion of the ECOFRAM draft report, EPA conducted an "Aquatic Peer Input Workshop" on June 22-23, 1999. One consensus view in the reviewer comments was that the models used for exposure assessment (PRZM/EXAMS *inter alia*) need validation. These validation studies should include field evaluation of both structural and parameter (scenario) reliability and performance characteristics. Case study examples of the proper use of the models in regulatory analysis should be developed as well. The goal of standardized and transparent scenarios for a variety of physiographic regions, crops, and aquatic ecosystem types was encouraged, with an ultimate aim of developing a complete database suitable for systematic, regular use in pesticide risk assessments.

As part of the process of further developing and implementing probabilistic risk assessment approaches, OPP/EFED executed a case study including both deterministic and probabilistic risk assessments. The aquatic assessment was limited to four crops (corn, cotton, potatoes, and grapes) for a single example chemical and product formulation [12]. In exploring the statistical properties of outputs from the PRZM/EXAMS linked exposure models, it was observed that no single distribution family consistently fit the output data and the quality of fit varied widely among scenarios. A 2-D Monte Carlo model was

therefore developed using an empirical distribution function to represent exposures, in preference to a fitted theoretical distribution. This study was presented to OPP's FIFRA Scientific Advisory Panel (SAP) in March of 2001. In its critique [13], the SAP strongly endorsed the direction taken by OPP, and "the Panel concluded that field verification (for effects and chemical fate) of model predictions is very important and needs to be conducted."

1. The purpose and scope of the assessment should be clearly articulated in a "problem formulation" section that includes a full discussion of any highly exposed or highly susceptible subpopulations [that have been] evaluated. ... The questions the assessment attempts to answer are to be discussed and the assessment endpoints are to be well defined.

2. The methods used for the analysis (including all models used, all data upon which the assessment is based, and all assumptions that have a significant impact upon the results) are to be documented and easily located in the report. This documentation is to include a discussion of the degree to which the data used are representative of the population under study. Also, this documentation is to include the names of the models and software used to generate the analysis. Sufficient information is to be provided to allow the results of the analysis to be independently reproduced.

3. The results of sensitivity analyses are to be presented and discussed in the report. Probabilistic techniques should be applied to the compounds, pathways, and factors of importance to the assessment, as determined by sensitivity analyses or other basic requirements of the assessment.

4. The presence or absence of moderate to strong correlations or dependencies between the input variables is to be discussed and accounted for in the analysis, along with the effects these have on the output distribution.

5. Information for each input and output distribution is to be provided in the report. This includes tabular and graphical representations of the distributions (e.g., probability density function and cumulative distribution function plots) that indicate the location of any point estimates of interest (e.g., mean, median, 95^{th} percentile). The selection of distributions is to be explained and justified. For both the input and output distributions, variability and uncertainty are to be differentiated where possible.

6. The numerical stability of the central tendency and the higher end (i.e., tail) of the output distributions are to be presented and discussed.

7. Calculations of exposures and risks using deterministic (e.g., point estimate) methods are to be reported if possible. Providing these values will allow comparisons between the probabilistic analysis and past or screening level risk assessments. Further, deterministic estimates may be used to answer scenario specific questions and to facilitate risk communication. When comparisons are made, it is important to explain the similarities and differences in the underlying data, assumptions, and models.

8. Since fixed exposure assumptions (e.g., exposure duration, body weight) are sometimes embedded in the toxicity metrics (e.g., Reference Doses,...[96-hour LC_{50}]), the exposure estimates from the probabilistic output distribution are to be aligned with the toxicity metric.

Text Box 2. EPA Policy: "Conditions for Acceptance" of Probabilistic Risk Assessments [10]

Modeling Issues and Technique

Pesticide Exposure Modeling

Models serve a variety of purposes in the scientific enterprise. In the most general terms, a "model" is a representation of some aspect of physical reality intended to enhance or express our understanding of that reality. Thus, physical models may serve as tools for studying the hydraulics of river systems, architectural models display the visual esthetics of a proposed structure, experimental models facilitate medical investigation of the causes of disease, statistical models help interpret the results of experiments, and mathematical process models express phenomena in symbolic formalisms. The goal of encapsulating physical phenomena in the mathematics of underlying process has its roots in antiquity, as for example in the work of Pythagoras (fl. 530 B.C.E.) on the integer physics of vibrating strings – the foundation of Western music. Models based in underlying process are almost universally held to be more reliable guides to action than are statistical extensions of observed trends or tendencies, an idea often expressed as "correlation is not causation." Mathematical process models occupy pride of place in the sciences because they can be readily manipulated to derive the consequences of the understanding imbedded in their structure. Examination of these consequences then serves as a test of the adequacy of the perceptions of reality underlying the model. The testing process is beset, however, with intractable epistemological problems with troublesome ethical implications for the use of models to inform public policy and regulatory decisions [14].

Models are notoriously easy to create and notoriously difficult to evaluate. Because the creation of models is intrinsic to science, the modeling enterprise has resisted attempts at standardization on the (quite legitimate) grounds that objectives and methods are unique to each discipline. In the context of "regulatory models," in which mathematical constructs have been encapsulated in computer codes to assist with regulatory decisions and rule-makings, there are strong economic and political reasons to cast doubt upon the reliability of these tools–at least when the conclusions are unpalatable. Regulatory decisions often embody an attempt to mediate among competing interests, and so are often unpalatable to some. Hostile scrutiny of models underlying regulatory decisions is thus the norm; the value of standard methods for evaluating regulatory models is apparent.

The questions routinely asked of technical staff by EPA pesticide regulatory officials include [15]:

> "What models did we use? Have they been validated? Are they widely accepted and scientifically sound?" and
>
> "How predictive and confident are we in using them?"

These simple questions raise deep issues of the basis and reliability of knowledge, the social aspects of technique, and the limits of forecasting in the face of incomplete information. Can a qualitative epistemological question ever be answered in the simple affirmative? Quantitative analysis can contribute to discussion of the weight and kinds of evidence required to support a decision, but cannot of itself create a value structure to support regulatory judgement. All models are fundamentally linguistic constructs, i.e., symbolic representations of an external physical reality. Perception of phenomena, modes and character of data collections, mathematical and logical inference, and the nature of available decisions and decision processes are inextricably intertwined. Because language provides the vehicle of communication and creates much of the controversy over model validation, clarity of language is the first task: in Oreskes' [16] phrase, "calling a model validated does not mean it is valid."

Concepts of Model Building and Model Testing

The terms to be understood include *validation*, *verification*, and *calibration*. In brief, *validation*, as used in the vernacular exhibited above, obviously pertains to suitability and reliability for the task at hand (from its Latin root *valere*–to be strong[3]). Its first meaning is one of having legal force, as in a *valid* automobile license plate. It has also, however, seen extensive use in scientific contexts as a term for experimental confirmation of theory, or as a term for model testing, e.g., "a process of obtaining assurance that a model is a correct representation of the process or systems for which it is intended" [17]. *Verification*, with roots in Latin *verus* (truth), speaks more immediately to the issue of "truth," although here too many of its

[3] Lexicographical discussion based on *Webster's Third New International Dictionary* (Unabridged), G. & C. Merriam Company, Springfield MA (1971).

senses are of the courtroom rather than of the state of our knowledge. *Verification* has sometimes been restricted to issues of the internal integrity of models [18], but has also been used in ways synonymous with some form of external *validation* [e.g., 19]. The distinction in common usage remains clear, however: a police officer may have a need to *verify* that a license plate is *valid*. Finally, *calibration* carries an idea that parameters of a model can be adjusted to "ascertain the proper correction factors" via comparing model outputs with observed data–rather as though a model were a physical measuring device in need of tuning to achieve its best performance. Natural language derives it power from its allusive force, so the bald assertion that a model is "valid" or has "been validated" inevitably suggests that its ability to make accurate predictions has been established. In actuality, such an assertion may mean almost anything, or nothing at all, and may well be met with the counter that "[v]erification and validation of numerical models of natural systems is impossible" [20]. This last idea is not merely provocative: the limits of empirical scientific knowledge as reliable guides to action are too easily forgotten in our technocratic era. Unfortunately, the incompleteness of scientific knowledge has been used as Luddite artillery by industrialist and environmentalist alike to attack inconvenient regulatory decisions. The *validity* of pesticide exposure models, and what is meant by it, is thus a matter deserving of close and detailed attention.

The difference between *interpolation* and *extrapolation* is also germane to what follows. Broadly, *interpolation* is the process of estimating intermediate values that fall within the range of prior observations, and *extrapolation* is the process of extending estimation beyond the range of prior observations. Interpolation often can rely on straight-forward statistical description of a dataset. There are pitfalls even in simple descriptive models, however.

In **Figure 1** the pattern of y as f(x) is described with a straight-line relationship, a curvilinear relationship, and two high-precision equations (empirical models!) providing an exact fit to all the observations (ignoring, as usual, the possibility of error in the measurements). The straight line is the simplest model of these data and, if Ockham's razor in its naive form ("simpler is better") is applied, it is the best. Noticing that the residuals exhibit some systematic structure (the midrange values tend to fall well above the line; the end members are all below) usually is sufficient justification for imposing a curvilinear relationship–in the name of being more "faithful to the observed data." William of Ockham's[4] stricture has not, incidentally, been violated, for his injunction was not "simpler is better," but rather "plurality should not be assumed without necessity" (or more precisely "non sunt multiplicanda entia praeter necessitatem"). Necessity is clearly a matter of judgement, so Ockham's razor is not quite the reliable guide to virtuous model-building sometimes alleged, and "keep it simple, stupid" is at best a grotesque modern parody of his idea.

The pitfall of an obsessive "faithfulness to the data" is also illustrated in **Figure 1**: two functions have been fitted to the dataset, each of which passes through every observed point with absolute fidelity. Note, however, that they disagree at every *interpolation* point, illustrating a point to which we shall return: there is a multiplicity of models (or, equivalently, parameterizations of model structures) equally able to represent any dataset. In this case, each of these two models can claim high fidelity (in both accuracy and precision) as a result of calibration. Which would be the better for prediction? Simply using the mean and variance of the observations would be preferable to a policy of despair in which every possible parameter set and descriptive equation is given equal weight in estimating the "uncertainty" of interpolated predicted values. Thus: simple numerical testing of models is an incomplete and error-prone (albeit necessary) guide to reliability, even when no more than a simple *interpolation* among observations is all that is desired. *Extrapolation*, the extending of estimation beyond the range of observation, invites another flavor of disaster altogether.

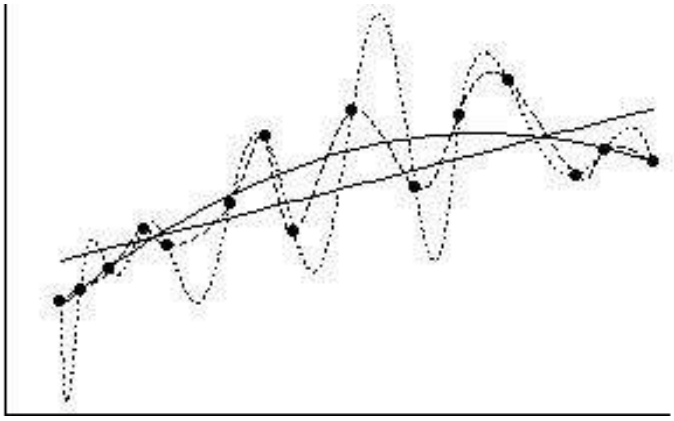

Figure 1. The dangers of interpolation (adapted from [21]).

The dangers of *extrapolation* [21] are illustrated in **Figure 2**. The two functions illustrated describe the observations equally well, and for interpolation purposes there is little need to choose between them. Once beyond the range of observation, however, their behavior is completely different. The choice, if interpolation will not suffice, must be made on grounds wholly external to the data itself. If, for example, the matter at hand is one of crop yields in response to soil amendments, Liebig's "Law of the Minimum" suggests the dotted line as the more reliable guide. If the data represent enzyme activity as a function of pH, the dashed line is the more plausible. The empiricist ideal of "letting the data speak for itself" cannot answer the need.

[4] b. *c.* 1285, Ockham, Surrey?, England; d. 1347/49, Munich, Bavaria (now in Germany). Franciscan philosopher, theologian, and political writer, a late scholastic thinker regarded as the founder of a form of nominalism and a pioneer of modern scientific epistemology. Marilyn McCord Adams, *William Ockham*, 2 vol. (1987), discusses in detail his thinking on a variety of complex topics; an informative summary of his life and thinking can be found in the Internet Encyclopedia of Philosophy (http://www.utm.edu/research/iep)

Empirical Science and Numerical Extrapolation Models

The principal distinction here is that of a descriptive, or correlative, as opposed to an explanatory, model [21] (**Figure 3**). The function of the correlative model, often developed by statistical analysis of experimental or observational data, is to provide some "adequate" function describing the data, which is then used as a predictive tool. The process is one of collecting relevant data, passing the dataset through some statistical machinery, and then using the resulting parameterized mathematical function for prediction as in **Figure 1**. The predictions can form a basis for additional investigations, from which additional data can be collected to improve the calibration of the model. The left side of **Figure 3** illustrates the process.

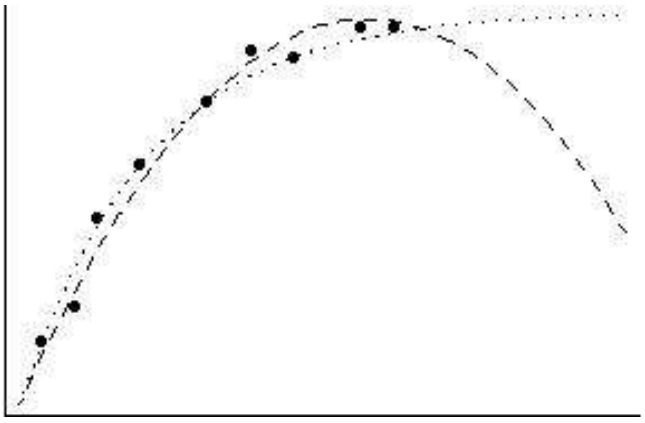

Figure 2. The dangers of extrapolation (from [21]).

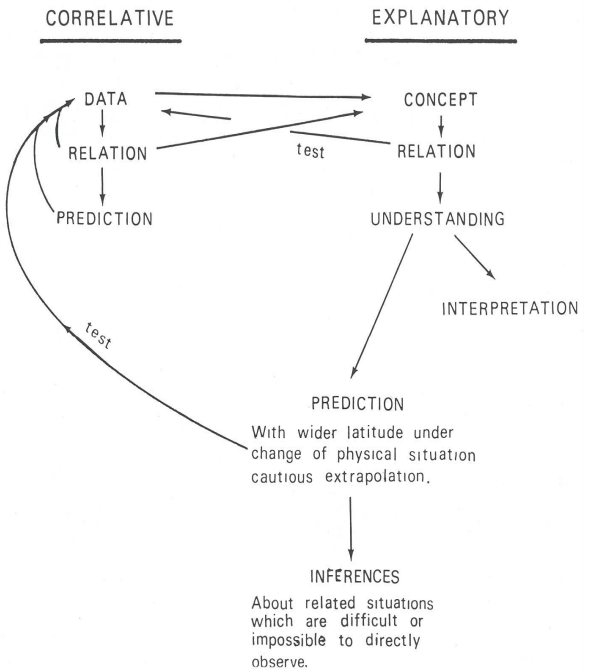

Figure 3. The model-building process [21]. Scientific models begin from empirical data. Conceptual models may be formulated directly from the data, or by inference from descriptive correlation analysis.

The right side of **Figure 3** illustrates explanatory modeling, as contrasted with correlative models, in the language of [21]. The useful distinction here is not one of "theoretical" vs. "empirical" models, a dichotomy that is little more than a politically charged relict of an ancient dispute between Platonists and Aristoteleans. In both cases, scientific models begin from a set of empirical data collected because judged to be relevant to some phenomenon of interest. The construction of an explanatory model is driven less by the desire to describe the dataset than by an ambition to discover a physical mechanism underlying the observed behavior. This conceptual model is then captured by a mathematical relationship, and tested by examining its ability to describe the original data. To the extent it succeeds, understanding of the phenomenon is increased, and similar events are amenable to interpretation in terms of similar physical causes. To the extent it fails, revision of the mathematics to achieve a better fit is not an acceptable strategy: the physical reasons for the failure must be analyzed.

Prediction now becomes possible beyond the range of prior observation because the mathematics of the explanatory model is not so tightly tied to the details of the empirical dataset from which the analysis began. These new predictions can be tested against elements of the original data, or by collecting additional data for a direct test of the extrapolations. As the collection of successful ("confirmatory") predictions accumulates, confidence in the model increases, and its reliability for inference concerning related situations becomes established.

A major advantage of explanatory models over correlative models (or, in politicized terms, "theoretical" models over "empirical" models; "process" models over "statistical" models) lies in the explanatory models' reduced structural uncertainty. Explanatory models offer better capabilities for the extrapolations required to support policy decisions. These capabilities flow from the organic process understanding built into the fundamental structure of the model. In contrast, an "empirical" model confronted with new data may require an entirely new structure to maintain its "curve-fitting" ability, and may thus radically alter even its interpolated predictions.

This tale of the social history of a model derives it rhetorical strength from its echoes of the standard "hypothetico-deductive" model of the scientific enterprise at large [22, 23]. In this view, science proceeds from triumph to triumph by forming precise hypotheses regarding physical phenomena, constructing critical experimental tests, and thus from incontrovertible rejection of hypothesis proceeding to the next step in the long accretion of scientific wisdom. It has been suggested, indeed, that any scientific idea not thus subject to "Popperian falsification" is not worthy of the name of science at all, being presumably mere philately or mysticism. One of the more important limitations of Popper's ideas is their inability to recognize the importance of the interpretive branch of **Figure 3**, as for example in the guidance of medical research provided by Darwinian biology. Another weakness of this idealization arises from the

observation that scientific theories are seldom abandoned on the basis of a single (or multiplicitous) contrary observation, but are patched and re-patched until the absurdity of the situation forces a general reorganization of the underlying conceptual model–a "paradigm shift," in the much-abused phrase [24]. Just as a single falsification does not warrant the wholesale abandonment of a model, however, neither does a single (or multiplicitous) successful prediction warrant any but tentative, conditional, and skeptical confidence. Still, prediction is a necessary precondition to rational regulation, and governments are not likely to revert to the methods of the ancient Roman haruspex. Understanding the nature of principled objections to the use of models for policy formulation, and their limitations and appropriate use, is critical for responsible analysis in support of regulatory activities.

When models are used to guide regulatory decisions, the demand that the models' reliability as a guide be assessed is eminently reasonable. Because society increasingly demands infallibility of public officials, that regulatory decisions guarantee a risk-free environment, and that the economic costs of regulation be vanishingly small, Agency decision-makers pressure technical staff to guarantee the "validity" of decision tools. A fear then arises that Gresham's Law (bad currency drives out good) will virtually guarantee that models which are not claimed to be "valid" will be supplanted by models of less scrupulous patrimony. Because model "uncertainty" and its quantification are increasingly valued, however, the bald affirmation that a model has been validated and verified is increasingly met with a healthy skepticism. Issues of model validation are dealt with more completely beginning on page 13.

Variability and Uncertainty

Variability and uncertainty are separated by their occupancy of different probability spaces, and by their ramifications for decision-making [9]. Variability refers to intrinsic heterogeneity in a quantity. It cannot be diminished by additional study, but only better characterized. Rainfall is a good example of a variable quantity of importance for pesticide exposure. Rainfall varies from place to place, and over time at any given place. Rainfall can be characterized by its intensity, duration, areal extent, total volume, etc., depending on the purposes of the analysis. Models of pesticide export from treated areas can be developed on the basis of annual rainfall/runoff volumes, as averages across physiographic regions, or as responses to specific rainfall events. Individual rainfall events may be characterized by 15-minute, hourly, daily, or annual totals. Model development is often constrained by the unavailability of suitably detailed input data to drive model responses to events.

Uncertainty, by contrast, can be reduced by additional study; it expresses a lack of knowledge of the true value of a quantity (which may also be masking variability). Uncertainty in pesticide risk assessment may arise from several sources. Imperfect measurements of chemical properties (e.g., transformation rate constants) are inevitable in the laboratory investigations underpinning registration studies. Chemical parameter uncertainty develops from random and systematic errors in measurement that can be adequately characterized using conventional statistical technique.

Exposure models are subject to structural (inadequate process algorithms) and parameter (inadequate parameter measurement) uncertainty. These often interact. For example, in characterizing the ability of rainfall to penetrate the soil, heterogeneity across an entire planted field may be summarized in a single infiltration parameter, which then becomes exceedingly difficult to measure during site investigations. The predictive reliability of models can be adequately characterized only by means of validation studies featuring direct comparison of model predictions with independent observations from empirical reality.

Additional uncertainties arise from spatial and temporal aggregations and approximations used to construct "scenarios" – the definitions of watershed geography and agroecosystem properties used to represent the treated landscape. Uncertainty also arises from the assumption that specimen "surrogate" systems can provide an adequate measure of protection for all protected endpoints (e.g., the use of farm ponds as a surrogate for all potentially affected aquatic ecosystems [13], or mallard ducks as a surrogate for all exposed waterfowl). Scenario and surrogate uncertainties are in large measure matters of regulatory experience and policy development.

In their ramifications, uncertainty "forces decision-makers to judge how probable it is that risks will be overestimated or underestimated for every member of the exposed population, whereas variability forces them to cope with the certainty that different individuals will be subjected to risks both above and below any reference point one chooses" [9].

The distinction between uncertainty and variability is sometimes difficult to maintain. The frequentist view of probability carries an implied symmetry between frequency of occurrence and probability of detection. For example, if there is a 10^{-3} chance that any given body of water is contaminated, then inspecting water samples from 1,000 water bodies would presumably locate the contaminated member of the set. Or, put another way, if a water body is selected at random, there is only a 0.001 chance of detecting contamination. This symmetry can lead to some confusion in risk communication: In a population of 200 million individuals, an individual risk of 10^{-6} is negligible for each single person, but the statement seems to imply that 200 individuals are at serious risk. Similarly, variability can contribute to uncertainty: when a quantity varies by several orders of magnitude, a precise estimate of the true mean can require a dauntingly large data set.

Probability distributions are used to quantify both variability and uncertainty, but the interpretation given the distributions differs. The concepts can perhaps be separated by reserving the term "frequency distribution" to describe variability, and "probability distribution" to represent uncertainty [25]. Distributions for

variable quantities thus represent the relative frequencies of values drawn from specified intervals, and distributions of uncertain quantities represent the degree of belief or subjective probability (colored by measurement error, etc.) that a specified value falls within a specified interval [26]. Uncertainty regarding variability can be viewed as probability regarding frequency [27].

It should be acknowledged, then, that variability and uncertainty are to some degree inextricably intertwined. For example, statistical summaries may capture variability in a few distributional parameters, sacrificing spatial or temporal precision for the sake of reducing a welter of detail to manageable proportions. Again, regularly spaced point measurements of the concentration of pesticides in watercourses sacrifice temporal detail to economic restraints, leaving uncertainty as to the true duration of episodes of contamination driven by rainfall events. Still, the motivation for separating natural variability from uncertainty is to clarify the contributions to decision processes of irreducibly stochastic phenomena, e.g., rainfall variability, as against the contribution of more readily remediable measurement errors and knowledge gaps. Reductions in uncertainty, by, for example, improving the precision and detail of laboratory investigations of chemical properties, or through field investigations of the efficacy of mitigation techniques, can improve the reliability of regulatory decisions. In contrast, the irreducible variability of natural phenomena will always introduce unwelcome uncertainty into regulatory decision-making.

Dealing with Variability and Uncertainty: The Monte Carlo Method

The "Monte Carlo method" was developed at Los Alamos National Laboratory to solve problems in weapons design stemming from the need to combine "stochastic and deterministic flows" [28] in the study of the "interaction of high energy neutrons with heavy nuclei" [29]. In this problem, the path of any particle and its sequence of interactions with other particles is physically determined by its initial velocity. The initial velocity of a particle cannot, however, be precisely predicted. In studying this problem, it developed that "the practical procedure is to produce a large number of examples ... and then to examine the relative proportion[s]" of each of the potential outcomes, thus producing a frequency distribution predictive of the probable behavior of the ensemble.

Monte Carlo methods take advantage of the ability of computers to produce random numbers drawn from a uniform distribution, which are then translated into the actual distributions of the stochastic input variables. The method was concisely defined by Metropolis and Ulam [29]: "Once a uniformly distributed random set is available, sets with a prescribed probability distribution $f(x)$ can be obtained from it by first drawing from a uniform uncorrelated distribution, and then using, instead of the number x which was drawn, another value $y=g(x)$ where $g(x)$ was computed in advance so that the values y possess the distribution $f(x)$." The name of the method is derived from games of chance that play out according to prescribed rules following an initial random event (e.g., randomization of card order, roll of the dice, etc.). In discussing the method, Metropolis and Ulam [29] noted that practical applications must, in the specification of the functions $g(x)$, allow for covariance (correlations among input variables) to avoid physically impossible combinations of parameters.

In Monte Carlo simulation studies, the analyst must "perform a finite number of experiments" that translate input variability and uncertainty into the output distributions of interest. The best way to specify the appropriate "finite number" is never entirely obvious. If too few iterations are performed, the output distributions are so contaminated by sampling errors as to be unreliable, especially in the distribution tails that are usually of greatest interest for risk assessment. If too many iterations are employed, the analysis becomes computationally burdensome and inappropriate for routine use. One solution, essentially Bayesian in outlook, is to track the statistical properties of the output distributions as the simulation proceeds and terminate the process when these properties stabilize. This translates the difficulty into one of defining adequate stability, which is at least more tractable than that of *a priori* defining the number of "experiments" to be executed on the computer.

Several means of sampling the input distributions $f(x)$ are commonly used; these include simple random sampling, "trace matching," and Latin Hypercube Sampling (LHS). Simple random sampling is usually deprecated because of its inefficiency, i.e., an excessively large number of simulations is required to achieve stable output distributions. In "trace matching," only actual observed values of the input variables are employed. This has the advantage of not requiring that the input data be matched to a distribution for sampling. It thus eliminates the sometimes contentious process of deciding on proper representation of the tails of the input distributions and the proper means of interpolating among observations. It has the parallel disadvantage, however, of potentially failing to capture extreme values and of being unnecessarily contaminated with gross measurement errors. Proper use of observational data for Monte Carlo input thus demands effective quality control procedures for database development, and some independent knowledge of the intrinsic variability of the natural processes underlying the data.

In Latin Hypercube Sampling (LHS), the input distribution is divided into n intervals of equal probability, where n is at a minimum the number of simulations to be run. Samples are drawn once, without replacement, from each of these intervals, thus ensuring efficient representation of the entire range of the distribution. Sampling within the intervals may be either entirely random or by explicit choice of the median of the interval, a technique known as "median LHS."

Interpretation and Presentation of the Results of Monte Carlo Analysis

Monte Carlo methods can be iterated during exposure analysis. An initial simulation using mean values of input parameters can identify dominant fate processes from an EXAMS model's standard sensitivity analysis, with subsequent application of probabilistic techniques to the governing parameters of those processes. When variable and uncertain parameters are collected during simulation and paired with output exposure metrics, multiple regression analysis can indicate the contribution of each input to the output variability [30].

In presenting the results of a Monte Carlo analysis, a "tiered" presentation style is recommended by EPA [11], in which the level of detail increases with each successive tier. For example, the first tier may be a one-page summary with graphs of the final risk metrics and a summary of the study, the second tier an executive summary, the third a detailed report. Documentation of model inputs is an important element of these reports. Reliable, coordinated synoptic observational datasets offer a number of advantages for exposure analysis: prior documentation relieves the individual analyst of the need to research environmental data for each analysis, simple trace matching obviates the need to fit the observations to an input distribution, the tails of the (empirical) distribution are not in dispute because they need not be invoked, and covariance among inputs is automatically accommodated. Uncertainty in chemical measurements should be treated as Normal distributions in almost all cases. Variability in laboratory data results from the accumulation of errors classically giving rise to the Normal distribution. Although an individual set of measurements may show statistically detectable skew or kurtosis, it is very likely that such results arise from sampling error rather than a genuinely non-Normal measurement process [31].

The outputs from a Monte Carlo exposure analysis can be presented as cumulative exceedence curves. In constructing a particular analysis, the exposure findings should be matched to the biological endpoint and the available toxicity metric. For example, for mortality of annually reproducing fishes the 96-hour LC_{50} can be compared to the largest average 4-day concentration for each year at a given site. For multiple sites, the 90^{th} percentiles of the 4-day events could be assembled. If chronic data for reproductive failure is available, 21-day average concentrations during the breeding period can be assembled for single sites over multiple years, or for some percentile of the single site distributions over multiple sites. These distributions

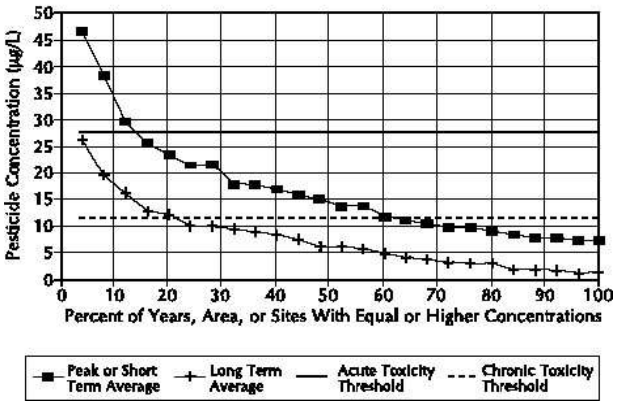

Figure 4. Cumulative exceedence curves constructed from single-site temporal data or multi-site data summaries.

can then be passed to a fuller probabilistic risk assessment incorporating uncertainty in the toxicity data, or compared directly to point estimates of toxicity as in **Figure 4**.

Full probability density function (pdf) and cumulative distribution function (cdf) graphs can convey different perspectives. These may also be constructed to match available toxicity metrics, including, for example, 24-h, 48-h, 96-h, 21-day, 90-day, breeding season or annual exposures for limnetic or benthic zones, for individual high-risk sites or across multiple sites representing the full suite of physiographic zones in the pesticide's use area. Point estimates and mean values from simpler analyses can be indicated on the these graphs for reference, as in **Figure 5** and **Figure 6**.

The pdf plot displays values of a random variable (concentration expressed in exposure metrics over short intervals) on the horizontal axis (abscissa) and relative frequency of occurrence or probability density on the vertical axis (ordinate). The pdf (**Figure 5**) is useful for displaying the relative probability of values, the most likely values (e.g., modes), the shape of the distribution (skew, kurtosis), and small changes in probability density.

A cumulative distribution function plots, on the ordinate, the probability that a value of the exposure metric (random variable) is less than a specific value on the abscissa. These plots (**Figure 6**) can display fractiles (including the median), probability intervals (including confidence intervals), stochastic dominance, and mixed, continuous, and discrete distributions.

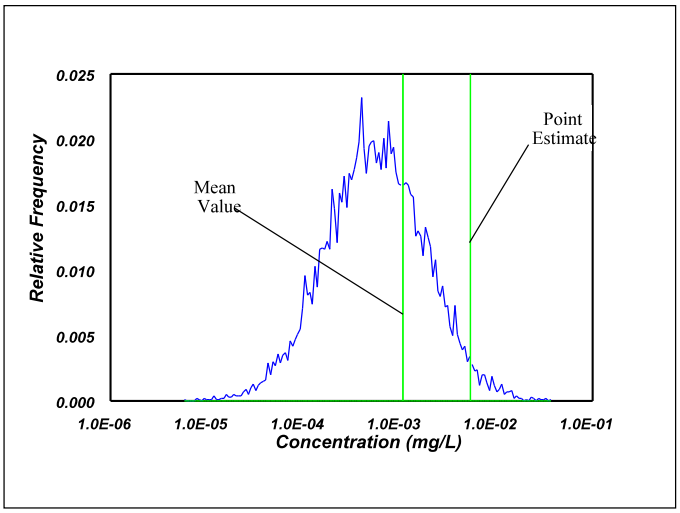

Figure 5. Example Monte Carlo estimate of a probability density function (pdf).

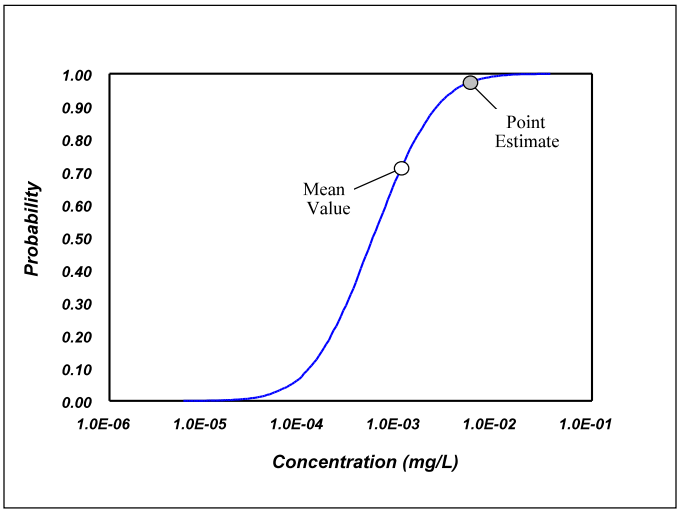

Figure 6. Example Monte Carlo estimate of a cumulative density function (cdf).

Validation, Verification, and Performance Measures

The Validation Problem

The idea that the "quality" of a model should be assayed is not entirely the same as the idea that models must be tested by comparing their predictions to independent empirical data. The "quality" of a model depends on (1) the quality of its underlying science, (2) the degree to which it has been shown to be "robust" (i.e., insensitive) in the face of violations of assumptions and approximations used in structuring the model, (3) "verification" studies confirming that the model indeed behaves according to its design specifications, and (4) demonstration studies of its parameter sensitivities and the relative performance of its internal constituent process models. These issues have been discussed in some detail in the context of the initial development of EXAMS [32], including, for example, tests of a simplified model of exchange at the benthic boundary layer, examination of the effect of neglecting the impact of sorption to suspended sediments on the photochemical absorption spectrum, and the structural and parameter sensitivity of volatilization models.

Although such studies are a necessary preliminary, studies of the performance of a model confronted with full-scale, independent empirical datasets reflective of the regulatory context are the only means of quantifying its "performance validity." Indeed, even models of very high quality (i.e., founded in well-established science, codes tested against an extensive set of limiting cases and parallel analytical solutions ("bench marked"), etc.) may fail when applied in a regulatory or policy development environment [33]. Such failures may arise, not only or even principally from failures in the underlying model, but from failures of the numerical representation to accurately solve the underlying equations in a complex field setting (for which there is now no equivalent analytical simplification), or from an inadequate representation of the environmental setting itself. In addition, there is clearly a difference between the ability of a model to describe a given situation ("history matching") by calibration of its parameters, and its ability to adequately represent new observations using the existing calibration, standard defaults, or generalized parameters. The latter has been termed a "post-audit" [34, 35] and provides the most fertile ground for generating a history of quantified measures of performance validity. Post-audit failures frequently expose the limitations of an initial calibration data set that failed to fully reflect underlying variability in environmental processes. More seriously, post-audit studies may reveal actual structural deficiencies, i.e., failures in the underlying conceptualization of the problem domain (see, for example, the case histories enumerated by Konikow and Bredehoeft [33]). These deficiencies are often then repaired, emphasizing that numerical models share some properties with scientific theory: they are often most useful as a means to enhance understanding and guide further work, and are liable to be patched and re-patched until they are supplanted by something entirely new. The task of quantifying prediction uncertainty is thus complicated by the felt obligation to use the results of a "validation test" to improve the structure of the model or, somewhat less usefully, to re-calibrate and thereby update knowledge of the parameter space able to generate acceptable history matches.

A further difficulty arises from the fact that, even absent re-calibration or re-structuring, successful empirical (post-audit) testing ("validation") cannot conclusively demonstrate that a model will have good predictive reliability when applied to novel situations [33, 36]. Thus, because "valid" is used by government agencies to signify "reliable for support of regulatory decision-making," Oreskes [16] has observed that "calling a model validated does not mean it is valid." To simply affirm that a model has been verified or validated *ipso facto* implies that its truth has been demonstrated, and it therefore can serve as a reliable basis for regulatory decisions. The tendency of scientists to claim that *validation* and *verification* mean something else within the technical community is ingenuous. The terms have indeed been given a variety of precise and honest definitions in the course of technical discussions, e.g., "Verification is usually defined as ensuring that the model behaves (runs) as intended, and validation is usually defined as determining that an adequate agreement exists between the entity being modelled and the model *for its intended use*" (emphasis added) [37]. That said, it must be observed that resistance to direct translation of the terms *valid* and *verified* into public discourse has been futile. It is equally futile to demand that the terms be abandoned, however, because a refusal to answer the query "Have [the models] been validated?" [15] would be viewed as obfuscation. The answer rather lies in a serious response to the collateral questions posed: "Are they widely accepted and scientifically sound?" and "How predictive and confident are we in using them?" [15] which address issues of the technical basis of a model and the degree to which uncertainties in its predicted values can be quantified.

Why Can't Models be Proven?

Absolute proof of a proposition or assertion can *only* be accomplished within the confines of a closed logical system – geometry, mathematics, symbolic logic. Within the natural

sciences, truth is elusive and ultimately unattainable, for a variety of reasons. First and most importantly, any set of observations we make on nature comes about from the operation of many underlying processes. Thus, data collected to test a model, even when completely consistent with the model predictions, do not "prove" the model, they only fail to disprove it. When the data and the model disagree, it is usually difficult to pinpoint the cause, not least because the model contains many components any one of which may be the primary root of the problem. The model may utilize well-corroborated components that are relevant to underlying process mechanics but are overridden by governing processes at the differing temporal or spatial scale of the test situation [38]. Similarly, when the model and data agree, the presence of multiple components introduces the possibility that the agreement is wholly fortuitous, for errors in one component may have been offset by errors in another. Although modelers occasionally claim to find this a comfort, it is unlikely that errors offsetting one another in one situation would continue to oblige in another. In addition, numerical models in the natural sciences are "under-determined" in their parameters, i.e., there is a very large set of parameter values that will result in a given set of output values. It is thus impossible to obtain a unique set of calibrated parameters to match any given set of field observations.

Numerical model extrapolations are necessary to evaluate the behavior of new pesticides, or of new uses proposed for existing compounds. These predictions are, however, unavoidably contaminated by uncertainties intrinsic to both the modeling and the data-gathering process. In structuring a process model, the level of detail chosen for the model requires that some detail – spatial, temporal, or causal – be either summarized or omitted as irrelevant. For example, the amount of suspended matter in a water column may be portrayed as a single concentration value, of specified organic matter content, averaged over a full month, applying to the full length of a stream reach. This representation is clearly false: suspended matter usually contains a mixture of particle sizes, some of which will deposit from the moving water and some of which will remain suspended, the amount of sediment and its particle size distribution changing during the month as rain storms pass through the area, etc. Difficulties with testing this model arise from several sources. Sediment-associated chemical is represented as a single data element, presumably for the purposes of evaluating the removal of the chemical from ecotoxicological concern. When testing or parameterizing the model, however, the stream can only be sampled at specific places and times, with an assumption that the sampling program can adequately represent the "true" mean values required for the model (even leaving aside the underlying question of whether this model adequately represents toxicological bioavailability). Uncertainty thus adheres to the data used to parameterize the model, the data used to test it, and to the approximation originally used to construct the model.

Increasing the detail and complexity of the model does not solve the problem. The data collection effort now becomes more complex, with a dependency on more complex instrumentation and analyses with their own uncertainties. The new model structure is conditioned (as envisioned in this example) on the selection among competing multi-phase sorption models developed under different laboratory conditions with a different set of experimental uncertainties. An increase in the knowledge embedded in the model carries substantial costs.

Although a more complex model incorporating additional knowledge of the system or additional spatial or temporal detail intuitively should provide more reliable predictive power, the necessary increase in its parameter complexity increases its level of under-determination. The number of sets of parameter values equally able to achieve a match to the observed history thus inevitably increases with model complexity. Because many of these parameter sets will lead to contradictory predictions when the model is confronted with new external data (e.g., a new chemical or environmental setting), a good ability to match prior experience may *not* be indicative of model reliability as a guide to good regulatory decisions [36]. Many of the important tests of a model must therefore be conducted in terms of critical tests of its constituent hypotheses. Other tests should be conducted at the boundaries of its use in current regulatory contexts in order to test its ability to provide useful information under novel conditions. Although model testing in particular field situations can serve to demonstrate a model's ability to reflect particular realities, only in the unlikely circumstance that this particular field setting were fully representative of the more general safety evaluation problem would the problem of establishing predictive model scenarios be resolved. The problem of "model validation," as conventionally phrased, may thus be an unreliable guide to the utility of a model for prediction and reliable regulatory guidance: once again, "calling a model validated does not mean it is valid" [16].

Testing the Performance Capabilities of Models

Despite the substantial intellectual resources devoted in recent decades to constructing quantitative frameworks for testing operational models [17-19, 37, 39-54], specific simple procedures for appropriately quantifying model performance remain incompletely developed. Quantitative assessment is too frequently replaced with subjective evaluations consisting of little more than a time-series plot of point observations and model-produced continuous lines, coupled with qualitative statements as to the adequacy of the "fit." The problem of subjective validity criteria is especially acute in the case of models designed for hazard and risk analysis of toxic chemicals, because of a continual conflict in the model user's, as against the model builder's, perception of the risk and decision structures surrounding validation studies.

How Can Prediction Uncertainty be Quantified?

Two (related) approaches to establishing the intrinsic accuracy and precision of a model are feasible: a descriptive approach, in which measures of model performance are accumulated to give a continually improving picture of model reliability, or a

hypothesis-testing approach in which the model's ability to meet pre-established performance standards is tested by some appropriate statistic. In either case, a Bayesian perspective is required: no single test can serve to unambiguously validate or invalidate a model. A series of tests can serve, in the aggregate, to establish confidence or fully discredit the model. Clearly, classical Bayesian judgement is required: the point at which enough testing has been done to declare the process complete is an entirely social decision. The descriptive approach suffers from an inability to evaluate the significance of any individual test because the performance criteria are unspecified and there is, therefore, a daunting array of possible performance reports [19]. The hypothesis-testing approach suffers from the need to establish performance criteria in advance to give a yardstick for evaluating the model. The descriptive approach can, however, be construed as an extension of the hypothesis-testing approach, and the hypothesis-testing approach also can serve to frame the discussion in terms of the risks involved in using models to guide policy and regulatory decision-making.

Risk and Decision Analysis in Model Testing

In all analyses, there is a risk that the analyst will infer the wrong decision from the test data. There are two varieties of such errors: in Type I, or *alpha* (α) error, the analyst mistakenly accepts the alternate hypothesis (H_a) when the null hypothesis (H_0) is in fact true. In Type II, or *beta* (β) error, the analyst accepts the null hypothesis as being true when the alternate is in fact true. The most cogent analysis of the role of risk in inference and decision was developed by Blaise Pascal (1623-1662, French mathematician, physicist, and philosopher). Pascal clarified the role of the *consequences* of wrong decisions in directing the analyst's attention to the appropriate choice of error controls, as in the example of **Table 1**. Here it is clearly most important to minimize the probability of a Type I (α) error (rejecting a true H_0), because of the significant negative consequences involved. Therefore, α, the probability of mistakenly rejecting a true H_0, should be made as small as possible, and β, the probability of a Type II error, can perhaps for most purposes be ignored.

R.A. Fisher, in his development of statistical practice for application to agricultural experimentation, understood that the tendency of the observer to desire a particular outcome is often a complicating factor in experimental design. For example, when the effect of a new fertilizer or pesticide on crop yields is under investigation, there is a natural tendency to want the new material to prove to be efficacious. This tendency to see positive results where none exist can be guarded against (**Table 2**) by posing H_0 as a "null" or no-difference hypothesis, and then making it difficult to reject (i.e., minimize the probability of an α error). In this instance H_0 is posed vs. a composite single-sided alternative that improvement exceeds some critical factor "δ," i.e., the material must produce a "substantial" improvement, and should it *damage* the crop it is of no further interest. Notice, in **Table 2**, that this device for institutionalizing intellectual honesty also allocates the more severe consequences of wrong decisions (ruin of individual and corporate reputations for probity and competence) to α errors. β errors generally merely lead to the question, "Are we sure this material should be abandoned?" – with the option of reserving judgement and extending field trials if the material is believed on other grounds to perhaps be efficacious after all. When this "no difference" null hypothesis approach to H_0 is used uncritically for testing the adequacy of ecotoxicological models, however, difficulties begin to arise. In what follows, "validation" will be used as a convenient synonym for "quantitatively testing the ability of a model to meet performance criteria," or, in the phrase of [37], "determining that an adequate agreement exists between the entity being modelled and the model for its intended use." The primary tasks include, first, establishing a quantitative definition of "adequate agreement," and then defining appropriate test methods.

Goals and Constraints of Performance Tests

The term "null hypothesis" (H_0) generally, although not of necessity, is used to refer to a condition of "no difference." Almost all validation studies have been constructed around a null hypothesis that "the model is valid," that is, that the mean values of the model predictions and of the observations on the prototype are not "significantly different" [40]; or that "model error is negligible" [43]. This approach has several inherent difficulties.

First, the observations on the prototype and the measurements of the parameters used to drive the model are both subject to error. Because the variance in the parameters propagates into the

Table 1. Probability matrix for hypothesis testing

If the truth is:	and the decision taken is to:	
	Believe	Not believe
H_0: God exists	Correct decision. Probability of this *correct* decision = $(1-\alpha)$. Consequence: Heaven	Analyst makes α error. Probability of this *wrong* decision = α. Consequence: Hell
vs. H_a: God does *not* exist	Analyst makes β error. Probability of this *wrong* decision = β. Consequence: unmerited faith	Correct decision. Probability of this *correct* decision = $(1-\beta)$. Consequence: atheism

Table 2. Decision matrix for agricultural efficacy experiment

If the truth in fact is:	and the decision that can be made is:	
	No improvement in yield	Improvement in yield
H_0: Treatment ineffective: no improvement in yield	Correct: P=(1-α) Consequence: discard faulty material	*Wrong*: α error (P=α) Consequence: tout faulty goods, ruin for all
vs. H_a: Improvement in yield of size δ.	*Wrong*: β error (P=β) Consequence: discard promising material	Correct: P=(1-β) Consequence: solve world food problem, prosper

model predictions, the paradox arises that a less precise model is better able to withstand validation tests. Notice, in **Figure 7**, that model M_2 is both less accurate, as well as less precise, than model M_1, but under conventional testing will be perceived as more "valid". For example, regression of experimental observations on model predictions has been suggested as a formal validation method [51]. In this case the validity of the model is judged against an ideal finding, phrased as a null hypothesis, of a slope of the regression line of 1 and an intercept of 0. This test suffers from the ambiguity depicted in **Figure 7**: the more scatter in the data, the larger the standard error of the slope, and the more difficult it is to reject the null hypothesis. Models with more scatter are thus less likely to be rejected [55]. (Data sets with few values or badly contaminated with measurement errors give rise to equivalent ambiguities.)

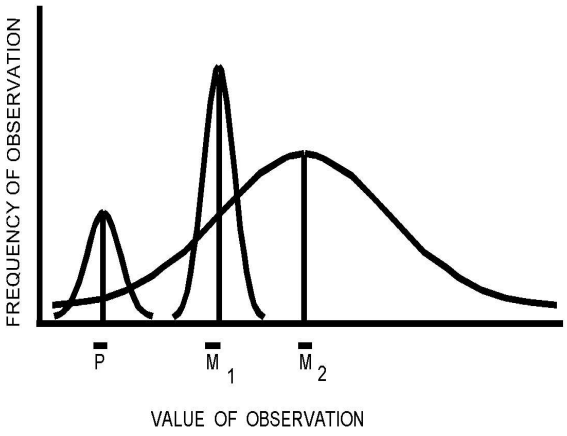

Figure 7. Distributions of observations on a prototype P and two models M_1 and M_2.

Secondly, when validation studies are posed so that the main hypothesis (H_0) is that there are no detectable differences between the model and the real-world "prototype," the primary hypothesis amounts to a statement that "the model is valid." As a result, the primary default statistical focus is on minimizing the risk of rejecting a true model (builder's risk), as a Type I (α) error.

The resulting decision matrix (**Table 3**) illustrates the fallacies of this approach. For a given sample size, in fact, increasing the stringency of the test statistic (by decreasing α) increases the likelihood of a β error. Regulatory scientists who must use environmental models are concerned, quite properly, almost exclusively with the dangers of using erroneous models. In this design, Type II (β) errors are the risk that they must perforce accept (model user's risk). Because Type II risks are usually not well-controlled, perceptive users of models are subject to continual Angst as to the reliability of their regulatory and evaluative tools.

A similar situation arises in experimental evaluations of the effects of pesticides on aquatic life. These tests have in some instances been conducted by dosing small ponds or "mesocosms" with the pesticide, comparing the dosed systems with undosed controls. When the efficacy model of **Table 2** is employed for analysis of the results (**Table 4**), the paradox arises that the most efficient way to certify the safety of a pesticide is to conduct an inferior experiment. The β error, which is poorly controlled or simply allowed to vary in response to sample size, contains what may be the more serious failure – allowing dangerous materials to slip through the safety evaluation and registration process undetected.

In some instances, validation studies have been designed to control user's risks by creating an especially demanding experimental frame that increases the likelihood of detecting false models [56]. The statistical power of such designs is always open to post-facto criticism, however, and they do not fully meet the need for wholly objective reporting on the performance and reliability of a simulation model. The problem, then, is that the focus of risk control in the usual run of validation studies is on the modeler's risk of rejecting a true model (Type I error), and the user's risk is ignored.

Testing methodologies can in fact be designed to control both Type I and Type II risks, while evaluating both accuracy and precision (bias and dispersion) of simulation models. These objectives can be met by reformulating the goals of validation studies to better conform with several principles of good statistical practice.

Table 3. Conventional (flawed) decision matrix for model validation studies

If the truth in fact is:	and the decision that can be made is:	
	No Difference	Model and Prototype Differ
H_0: No difference between "model" and "prototype" – model is therefore "valid"	Correct: P = (1-α) Consequence: accept model, use for decisions	*Wrong*: α error (P = α) Consequence: modeler's ruin: reject good model
vs. H_a: model and prototype substantially differ – model is fatally flawed	*Wrong*: β error (P = β) Consequence: user's ruin: faulty safety decisions	Correct: P = (1-β) Consequence: Better models must be developed

Table 4. Conventional (flawed) decision matrix for aquatic safety testing

If the truth in fact is:	and the decision that can be made is:	
	No Difference	Pesticide has an impact
H_0: No difference between "treatment" and "control" – pesticide is therefore safe	Correct: P = (1-α) Consequence: register safe pesticide, allow runoff	*Wrong*: α error (P = α) Consequence: ban materials that pose little danger
vs. H_a: treatment and control differ–pesticide is causing ecological harm (or benefit?)	*Wrong*: β error (P = β) Consequence: allow use of harmful material	Correct: P = (1-β) Consequence: ban dangerous materials; seek substitutes

Good Statistical Practice for Testing Models

First, when conducting a statistical evaluation of any experimental situation, good practice demands that the null hypothesis be phrased as the opposite of what the experimenter wishes to prove; the tests can then be constructed so as to protect against unconscious bias (stemming from the modeler's natural desire to have constructed a valid model) by making it difficult to reject H_0. In validating simulation models, then, the null hypothesis should always be formulated in terms of *H_0: the model is invalid*, vs. an alternative hypothesis *H_a: the model is valid*. (Note that this is the opposite of the usual practice in the field.) This has the advantage of re-assigning the user's risk to Type I error, and the modeler's risk to Type II error, and allows for a "no decision" option (continued model development) in addition to the usual "two-frame" approach to validation [57].

Second, whenever possible, statistical tests should be formulated so that H_0 is a "simple" (as opposed to a "composite") hypothesis that can be tested against a one-sided (composite) alternative, so that it is possible to specify α and β, the probabilities of committing Type I and Type II errors, and determine the minimum sample size needed to attain this degree of precision [58:262]. Parametric tests should be preferred over non-parametric techniques whenever possible: non-parametric tests often require fewer computations and assumptions about underlying distributions, but they also carry greater probabilities of Type II errors (rejecting a valid simulation model) for a given level of user's risk, a situation highly undesirable from the modeler's perspective!

Third, it must be recognized that both the model predictions and the observations on the prototype involve the same experimental unit, even though both are subject to (often independent) measurement errors. For this reason, only "paired sample" techniques are fully appropriate in most such situations. As in the case of non-parametric analyses, when inappropriate techniques (e.g., tests of differences between means) are applied, the frequency of Type II errors (modeler's risk) will tend to rise, in this case as a response to increasing variances. Finally, given a situation in which parametric tests can be applied, both user's and modeler's risks can be controlled at pre-determined levels during the design of the study by selection of an appropriate sample size [59].

A Methodology for Performance Testing (Validation) of Simulation Models

To apply these principles, it must first be recognized that objective test methods require objective criteria for judgment of model validity. These criteria usually must be based on considerations external to the model, that is, on the needs of the model user for accuracy and precision in model outputs. Chemical exposure models are usually used in a context of toxicological safety evaluations, in which, given the parallel uncertainty in toxicological data and inferences, a consistent

factor-of-two error is probably not excessive [60]. We can thus formulate an objective test of validity as follows:

> "Predicted values from the model must be within a factor of two of reality at least ninety-five percent of the time."

(The particular error factor and reliability criterion selected is not significant; the methodology would apply to a 25% error factor, or a factor-of-three error, a 99% adherence to the error criterion, or any other numerical specification of acceptable uncertainty and reliability in model predictions.) Paired sample techniques require that the validation focus on the point-to-point differences between the model and the prototype. The error in the prediction is a function of the difference between the model prediction "P" and the observation on the prototype "O". The observational pairs are usually taken over a period of time, leading to a series $\{P_1,O_1; P_2,O_2; ...\}$.

Some thought should be given to the effect of serial correlation and independence of the pairs on the significance of the test results. For example, when studying a series of contamination episodes in a watercourse, if the hydrology is relatively simple dissipation of the chemical may merely reflect flow characteristics. In this case, comparisons might be tailored to compare the predicted and observed integrated dose for each event, rather than comparing every measured concentration to its predicted pair. Complete point-by-point comparisons may, however, be entirely appropriate in a more complex hydrologic regime or a more complex water quality setting.

Predictions can be arrived at in several ways, for example:

- When the model parameters are also measured at times t_1, t_2, ..., the test could be called a "structural validation," that is, it tests the mathematical structure of the model.

- When the model is run with mean or nominal values of the parameters (degenerate random variables, i.e., of zero variance), the test could be termed a "parameter validation." It tests the ability of the model to function adequately when its input data are subject to larger errors and approximations.

- When the model has been designed for Monte Carlo sampling of the parameter space, parameter validation using degenerate random variables can be followed by Monte Carlo tests that incorporate stochastic effects into the model predictions. These tests could provide a statistical validation of models with explicit provision for stochasticity in their (environmental and/or chemical) input parameters.

Regardless of the type of validation test, paired-sample error measures allow for simultaneous stochastic effects on both P (the model predictions) and O (observations on the prototype). Errors in parameter estimations (expressed in P) and measurements (sampling errors in O) are both incorporated into the test procedure. As a result, the model with the larger variance (M_2 in **Figure 7**) loses its apparent advantage over its more precise cousin.

The first test that must be executed is a simple test for positive correlation between the P and O series. (Notice that regression techniques cannot be applied rigorously here because neither P nor O is known precisely.) If P and O are not positively correlated, then the mean value of O would presumably be a better model than is the simulation program. This test can be simply phrased as

$$H_0: \rho = 0 \text{ vs. } H_a: \rho > 0$$

where ρ is the population correlation coefficient, and testing is conducted at a significance level α of 0.01 in order to maximize protection from false models. (Notice that should it develop that $\rho<0$, the one-sided test will tend to support H_0 rather than H_a.) In designed validation studies, models are unlikely to fail this test because of the wide range of conditions typically selected for testing. In studies of real-world systems, failures are more likely; in any case correlation serves as a convenient screening test and confirmation that continued analysis is warranted. Although sensitivity analysis is of most value during model development, finding significant correlation in this test also serves to confirm that the model is sensitive to parameters that co-vary with the observed data in the current validation study.

Implementation of the validity criterion given above requires a measure of the observed differences D such that a situation of no difference between the P and O gives a value of zero, increasing to a critical value D^* at a factor-of-two difference. For parametric testing, the underlying distribution of the comparison function D should be Normal, and the sampled distribution of D in a specific test should be at least approximately normally distributed. A simple ratio is not symmetric on the interval of factor-of-two under- and over-prediction, and thus not a candidate. One candidate transformation is the logarithmic, such that $D = \log(P) - \log(O)$. This function has the desired properties:

at P = O,	$D = 0.0$
at P = 2 x O,	$D = +0.301$
at P = O/2,	$D = -0.301$

Thus, the maximum permissible bias (i.e., systematic over- or under-prediction), for a model with absolute precision (zero variance), would occur at $|D| = 0.301 = D^*$.

The validity criterion includes a precision constraint as well, here taken as a constraint that 95% of the errors must be within the limits of permissible bias. This requirement leads to a maximum permissible value for the standard deviation of D, in addition to its maximum mean value. For a completely unbiased model ($\mu_D = 0$), a 95% confidence interval on D may not exceed

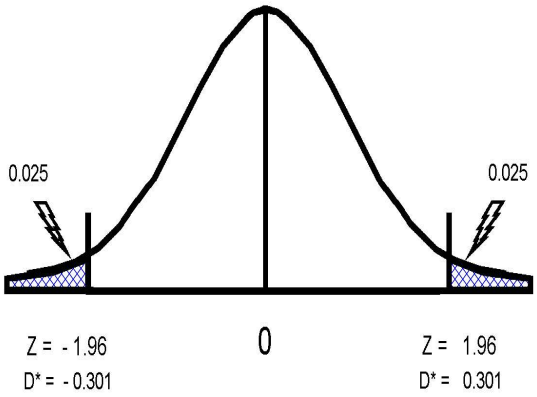

Figure 8. Maximum values of bias (D*) and dispersion (σ*) permitted of model meeting validation criteria.

the limits [-0.301,+0.301] (i.e., a factor of 2 (log(2)=0.301)) in order that the model be judged valid. For a normally distributed random variable, 95% of the area under the curve is encompassed by the interval -1.96 to +1.96 in *z*, where *z* is a random variable having the Standard Normal distribution N(*z*;0,1) (that is, *z* is normally distributed with mean zero and variance 1). The corresponding points on the underlying distribution of *D* (assuming it to be Normal) are the limits of permissible bias -0.301 and +0.301 (**Figure 8**). In order to compute the maximum permissible value for σ, the standard deviation of *D*, we need only observe that when *D* takes on any value X, the corresponding standardized normal random variable *z* assumes the value (X-μ)/σ. For an unbiased model (μ_D=0), the critical value of the standard deviation of *D* at the boundary of the region of acceptable variances (also D^*), say σ^*, is (**Figure 8**):

$$\sigma^* = D^*/1.96 = 0.301/1.96 = 0.154$$

This suggests that one test that must be performed during the validation is

H_0: σ = 0.154 vs. H_a: σ < 0.154

using, for example, a χ^2 test with α = 0.01. Note that the test is again phrased with a simple null hypothesis, a one-sided composite alternative, and is designed to preserve our interest in valid models by making it difficult to reject H_0. As with the test for correlation given above, if σ>0.154 the test will tend to support the null hypothesis and the model will fail validation.

Clearly a goodness-of-fit test of the supposition that *D* is normally distributed will also be required, which can be posed in the conventional way as

H_0: *D* is (approximately) normal, vs.

H_a: *D* is distributed in some other way.

This test can also be conducted via a χ^2 test, but in this case, the goal of protecting against false models suggests that it be somewhat *easier* to reject H_0. We can, for example, relax α to 0.10, in preference to α = 0.01, in order to continue to minimize model user's risks, in the sense that this makes it more likely that the premises upon which the analysis is constructed will be rejected. The statistic computed for this test is a value of a random variable whose distribution is approximately chi-square with (*p*-*q*-1) degrees of freedom (d.f.), where *p* is the number of cells and *q* is the number of parameters estimated from the data. As two parameters (μ and σ) must be estimated, a minimum of four cells (groups) is required to conduct this test (with only a single remaining degree of freedom). It is customary to use this test only when none of the expected frequencies is less than five [58:284]. This suggests that sample sizes should not be less than twenty in order to validate normality assumptions.

Alternative tests of normality are available for smaller sample sizes. The Shapiro-Wilk test [61] has on occasion been recommended by EPA as a "test of choice" for normality [11]. For this test, the *n* observations on *D* are ordered so that $X_i > X_{i-1}$, i=2,3,...n, and the test parameter *b* is calculated from

$$b = \sum_{i=1}^{n/2} (X_{n-i+1} - X_i) a_{in} ,$$

where the a_{in} are tabulated coefficients for the elements of the sum; the values of the a_{in} depend on the sample size *n*. W_n, the test statistic, is then calculated from $W_n = b^2/[(n-1)s^2]$, s^2 the sample variance of the X_i. If W_n is smaller than the critical value $W_{n,\alpha}$ the null hypothesis (of normality) is rejected. A description of the test and tables of the required coefficients and percentage points of the test statistic are available for $2 \le n \le 50$ in [62].

The method of constructing sample sizes for controlled levels of user's (α) and modeler's (β) risk [59] requires the specification of acceptable levels of both. In this instance, we will select both risk levels at 0.01. Clearly the appropriate values for these risks depend on the perceived social significance of the acceptance and use of false models, and of the rejection of true models; the methodology does not depend on the particular values selected. This methodology does possess the significant advantage, however, of making the risks explicit.

Combining bias and dispersion in the model performance criterion requires specification of the least effect that must be detectable in the analysis (δ). Because σ is unknown in paired-sample analyses, when calculating the minimum required sample size δ must be phrased as a function of standard deviation (S.D., here the S.D. of the error measure *D*). This value of the S.D. can be derived by considering the fact that all real models possess some bias, even if it is not detectable (i.e., not "statistically significant"). It should be emphasized that the simple presence of detectable bias is not of any particular interest in a validation

study because its presence can be conceded from the outset, and sufficient testing is certain ultimately to detect it. The points of concern are, first, whether the accuracy and precision of the model are sufficient to meet the validity criterion (here that 95% of predictions be within a factor of two of reality), and, second, the method of translating this criterion into an objective test that is based on the bias and dispersion of the error measure.

The smallest interval of interest (the least detectable effect) will occur when sufficient bias is present to allocate all violations of the validity criterion to one side or the other of the error distribution (**Figure 9**): this value is necessarily smaller than any instance in which the permitted 5% of predictions larger than the permissible (factor-of-two) standard are occurring as both under- and over-predictions. The left side of **Figure 9** depicts a case in which the model is producing predictions somewhat smaller than the observations on the prototype, that is, D is less than zero (the argument is symmetrical for $D>0$). Assuming that $\sigma < \sigma^*$ (which assumption will be validated in a separate test), we need concern ourselves only with the left-hand side of the distribution of D. The true value of D must be greater than $-D^*$ by an amount sufficient that only 5% of the total area under the curve is to the left of $-D^*$. In the standard normal distribution, this (single-sided) value corresponds to $z = -1.645$, and this is the interval "δ" which must be detectable in the validation study. (The final test will be posed in terms of single-sided alternatives, with selection of the appropriate test predicated on whether the estimated mean value of D is less than, or greater than, zero.) As was done above for determining the maximum permissible value of σ ($\sigma^* = 0.154$), the value of δ can be computed by taking advantage of the relationship between any normally distributed random variable X and the Standard Normal variate z: $z = (X-\mu)/\sigma$. As $z = 1.645$, and $(X-\mu)$ is the distance $D^*-\mu_D = \delta$, the difference that must be detectable in the study, substitution in this formula leads to $\delta = (D^*-\mu_D) = 1.645\ \sigma$.

The formula for computation of the required minimum sample size [59] is

$$N = (U_\alpha + U_\beta)^2\ \sigma^2/\delta^2$$

where U_α and U_β are the user's and modeler's risk points on the normal probability curve; here both have a (single-sided) value of $z_{0.01} = 2.326$ [59:325]. Substituting $\delta = 1.645\ \sigma$ yields a value of $N = 7.997$. As the true value of σ is actually unknown, this estimate of N is too low. When σ^2 is unknown, validity tests must be based on the (single-sided) t distribution rather than the Normal distribution, and the equation for N must be revised to read

$$N_t = (t_\alpha + t_\beta)^2\ \sigma^2/\delta^2$$

Taking N-1 = 6.997 degrees of freedom, the probability point on the (single-sided) t distribution for an 0.01 level of risk is 3.0 [59:326], and

$$N_t = (3 + 3)^2 / (1.645)^2 = 13.3$$

Therefore, the collection of 14 samples will control both user's and modeler's risk at a level of 0.01, for a normally distributed error measure having a standard deviation $\sigma < 0.154$. (Recall, however, that a minimum of 20 samples is required for a χ^2 test of the normality of the error measure; the Shapiro-Wilk test is available for N=14).

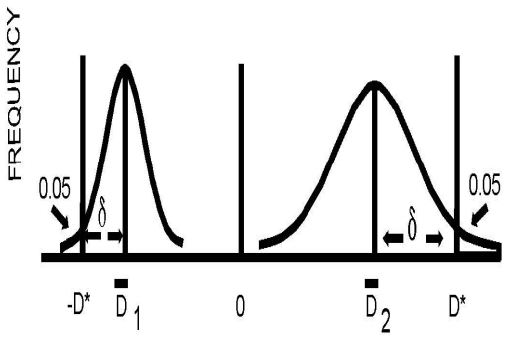

Figure 9. Definition sketch for formulation of objective validation tests.

The actual single-sided tests on the error measure can now be specified. These tests incorporate both the accuracy and the precision requirements of the validity criterion imposed on the model, with the selection of the appropriate test dependent on the sign of the sample mean of the error measure D (**Figure 9**). In order to phrase the final test as a simple (non-composite) null hypothesis, we will specify that the model will be rejected as invalid if it is at (or, by implication, beyond) the boundary of the validity criterion. This is a slightly more restrictive test than the original criterion, but its advantages in posing the null hypothesis for risk control greatly outweigh any loss of latitude in acceptable model performance. (The criterion was not initially specified in this form because of the convoluted language required: It now reads "the model must be closer to reality than a factor of two, at least slightly more than 95% of the time.") The appropriate hypotheses are, for $\bar{D} > 0$,

H_0: $\mu_D = (D^*-\delta)$, with the alternate H_a: $\mu_D < (D^*-\delta)$

and, for $\bar{D} < 0$,

H_0: $\mu_D = (-D^*+\delta)$, with the alternate H_a: $\mu_D > (-D^*+\delta)$

where δ, the least significant difference, is now computed from the sample standard deviation S of D, and the t distribution at a ("single-sided") probability point of 0.05 (if based on a validity criterion of 95% precision) with n-1 degrees of freedom:

$$\delta = \frac{t_{0.05,n-1}S}{\sqrt{n}}$$

Finally, the validity test is conducted, for $\bar{D} > 0$, by comparing the value from the single-sided t distribution ($t_c = t_{\alpha,n-1}$), α the level of user's risk, to the sample t statistic computed from

$$t = \frac{(\bar{D} - (D^* - \delta))\sqrt{n}}{s}$$ in which H_0 is rejected if $t < -t_c$.

When $\bar{D} < 0$, the sample t statistic is computed from

$$t = \frac{(\bar{D} - (-D^* + \delta))\sqrt{n}}{s}$$ and H_0 is rejected if $t > t_c$.

A Step-Wise Procedure for Performance Testing

The steps required for a validation study can now be enumerated:

Step 1: Select appropriate validity criteria and establish acceptable levels of user's and modeler's risk. (Except as otherwise noted, upon completion of Step 2 the level of user's risk selected in Step 1 should be used as the α risk for all subsequent analyses.) As here illustrated, one choice for ecotoxicological models is to specify that the model predictions and the observed data must differ by no more than a factor of two, at least 95% of the time, with the level of user's risk controlled at 1%, i.e., no more than one chance in a hundred of accepting an invalid model. In the example of DMDE photolysis in EXAMS (below), in which a constituent hypothesis of the model is being tested, the acceptable difference is set at a factor of 2.0×, at least 99% of the time, with a 1% user's risk of the model being falsely accepted as meeting the criterion.

Step 2: Develop an appropriate error measure for

H_0: the model is invalid–it cannot meet an acceptable range of accuracy and precision in the current context, vs.

H_a: the model is valid (acceptable), at least in this instance.

Step 3: Determine the minimum sample size requirements needed to control risks at acceptable levels and to provide sufficient data for tests of underlying assumptions.

Step 4: Collect point-for-point sample data on the model output and an experimental unit (prototype), producing a set of paired samples of simulation model predictions and observations on equivalent properties of the prototype system.

Step 5: Test the paired data for significant correlation; reject the model if they are not positively correlated.

Step 6: Compute the dataset for the error measure D, and test the hypothesis that D is approximately normal (via a χ^2 or other test at a level of significance suitably less restrictive than the α risk level in use to control user's risks). If the null hypothesis is rejected, the analysis is invalid and will at the least require reformulation.

Step 7: Test the null hypothesis

H_0: $\sigma = \sigma^*$, the model is unacceptably imprecise, vs. the one-sided composite alternative

H_a: $\sigma < \sigma^*$, the model precision is adequate, where σ^* is the critical maximum permissible variance derived from the validity criterion imposed in Step 1.

Step 8: Test the full validity of the simulation model by comparison of the computed value of the t statistic to the appropriate (α, ν) point on the t distribution, ν the degrees of freedom.

The decision matrix is given in **Table 5**; note the allocation of model user's risk to α error. If H_0 is rejected then the model can be concluded to be "not invalid" (at least in this instance) with certainty of at least (1-α), where α is the level of model user's risk selected in Step 1. In the DMDE photolysis example presented here, if $|t| > |t_c|$, the model can be accepted as having passed this particular performance test, with 99% certainty.

A Substantial Example: Photolysis of DMDE in EXAMS

The Exposure Analysis Modeling System EXAMS [4] contains algorithms for computing radiative transfer in the atmosphere as a function of location and climate, and for coupling the absorption spectrum and quantum yield of a synthetic chemical to computed spectral irradiance in order to arrive at an estimate of the rate of photochemical reactions in the aqueous environment. Zepp [personal communication; 63] has examined the clear-sky behavior of 1,1-bis(*p*-methoxy-phenyl)-2,2-dichloroethylene (DMDE) at solar noon, collecting 20 independent samples over the course of the calendar year. These experiments provide an opportunity to evaluate EXAMS' constituent hypotheses governing direct photolysis kinetics, within the context of the complete model system.

The EXAMS program calculates rate constants based on mean whole-day (i.e., over the 24-hour period) conditions on the day of the month when the solar declination results in the monthly mean value of the incoming extra-terrestrial irradiance [64:62]. For testing, EXAMS was loaded with the declination and radius vector of the Earth on the specific dates at which the observations O were taken. To form a commensurate dataset, the rate constant k computed by EXAMS must be cosine corrected for having been averaged over the course of a day, and then adjusted for day length:

k (noon value) = (daily value) $(\pi/2) \times$ (24/day length)

The parameter robustness of the model was investigated by loading the code with standard values of its environmental input parameters. Monthly mean values of stratospheric ozone were loaded automatically via specification of the latitude (33.94°) and longitude (-83.32°) of the Athens, Georgia (USA) laboratory at which the observations were collected. (See page 30 for a

Table 5. Decision matrix for unambiguous validation testing of environmental models

If the truth in fact is:	and the decision that can be made is:	
	Accept H_0–reject the model as "invalid"	Reject H_0–accept the model as "valid"
H_0: Dispersion or differences of model too large–model is thus "invalid"	Correct decision: P = (1-α) Consequence: continue model development	*Wrong*: α error (P = α) Consequence: accept bad model, make poor decisions
vs. H_a: model has acceptable precision and accuracy, and can be accepted as "valid"	*Wrong*: β error (P = β) Consequence: wasted effort fixing valid model	Correct decision: P = (1-β) Consequence: use "valid" model to improve safety

description of EXAMS' internal dataset, a compilation of data from the TOMS (Total Ozone Mapping Spectrometer) instrument flown on the Nimbus-7 spacecraft). EXAMS' input monthly atmospheric turbidities [65] were taken as a constant 2 km (the default monthly values) using a "Rural" atmospheric type. These data thus constitute a fairly rigorous test of the model's ability to withstand reasonable levels of error and approximation in its input parameters, and represent typical use in a regulatory environment in which the default data are routinely employed.

For testing a constituent process algorithm of EXAMS, the acceptable validity criterion, i.e., the uncertainty adhering to a point estimate from the model, was taken as "model and observation must differ by less than a factor of two, at least slightly better than 99% of the time, with 99% certainty." Putting **Step 1** in another way, passing the test will serve to certify, with less than 1% chance of error, that 99% of the estimates are within a factor of two of equivalent experimental observations. (Note that, so long as (kt), the product of the decay rate constant and time, is < -4, statements concerning half-lives are equivalent to within about 2% to statements concerning relative exposure concentrations.)

For **Step 2**, *D* is taken as log(P)-log(O) and *D**, the limit of acceptable difference, is log(2.0). The criterion of 99% of differences within 2.0× results in a maximum permissible standard deviation (σ^*)of $\log(2.0)/z_{0.01} = 0.301/2.576 = 0.1168$. (The area under the standard normal between -2.576 and +2.576 encloses 99% of the total.)

For **Step 3**, calculation of the minimum number of pairs required to achieve 0.01 α and β risks, the least detectable effect δ is based on the point on the normal probability curve for which 99% of the curve lies to the right of *D** (see **Figure 9**, here using 0.01, rather than 0.05, as the area outside *D**). The ("single-sided") value of $z_{0.01} = 2.326$, so $\delta = 2.326\sigma$. The minimum sample size to achieve a 1% Type I user's and Type II modeler's risk is $(3+3)^2/(2.326)^2 = 7$.

Step 4. Table 6 gives observed half-lives, EXAMS' predicted values, and the computed differences (*D* = log(P) - log (O)) .

Step 5. Correlation analysis to validate the assumption that the simulation model is a better predictor of the observations than is the sample mean. Correlation of the O and P datasets in **Table 6** yields a value of the sample correlation coefficient *r* of 0.93. In testing the null hypothesis (H_0: ρ = 0) vs. the one-sided alternative (H_a: ρ > 0) at an α=0.01 level of significance, the computed value of the *z* statistic must be greater than the single-sided $z_{0.01}$=2.326. Computing the *z* statistic associated with this value of *r* in the usual way [58:311] gives

$$z = \frac{\sqrt{n-3}}{2} \ln \frac{1+r}{1-r} = 6.62$$

The null hypothesis can thus be safely rejected, leading to the conclusion that the observations and model predictions are indeed positively correlated and the analysis may proceed.

Step 6. Test the (approximate) normality of the distribution of the error measure D. This step tests the null hypothesis

H_0: *D* is approximately normal,

vs. H_a: *D* is not normally distributed and the analysis must be reformulated.

To execute the χ^2 test, the sample mean and standard deviation are used to predict the frequency distribution of the observations *D* on the pairs. These expected values are then compared to the actual distribution of the sample differences via a χ^2 "goodness-of-fit" test. The test can be conducted by partitioning the observed data into the required minimum of four cells (to have at least one d.f.), each of which meets the requirement of expected frequency ≥5, by constructing four cells of equal frequency centered on the sample mean (-0.0032). The value of *z* in the Standard Normal (0.675) giving the 25% points was converted to matching intervals on the distribution of *D*, assuming *D* to be normally distributed with mean -0.0032 and standard deviation 0.0707. This has the effect of partitioning *D* into 4 cells with expected frequency of 5 values of *D* in each cell. The 20 observations in **Table 6** were then assigned to the matching cells in **Table 7**.

From the results shown in **Table 7**, the χ^2 statistic is computed in the usual way as the sum of the $(O-E)^2/E$, giving in this case a value of 0 (zero). This value is less than the (0.10, 1) critical value of χ^2 (2.70), so the null hypothesis that the distribution of *D* is approximately normal cannot be rejected. Execution of a Shapiro-Wilk test on the ordered set of *D* yields W=0.9659. As $0.9659 > W_{(20,0.10)}=0.920$, the null hypothesis that *D* has a normal distribution again cannot be rejected and the analysis may proceed.

Table 6. Observed and predicted EXAMS mid-day half-lives of DMDE at Athens, Georgia.

Date	Ozone (O_3)[1] cm NTP	Mid-day Half-lives (hours)		Difference metric (*D*)
		Observed	Model	
2 Jan	0.282	4.41	3.35	-0.1194
23 Jan	0.282	3.38	2.69	-0.0992
31 Jan	0.282	2.30	2.39	0.0167
19 Feb	0.296	1.67	1.81	0.0350
24 Apr	0.316	1.36	1.03	-0.1207
21 Jun	0.312	0.96	0.93	-0.0138
23 Jun	0.312	1.02	0.93	-0.0401
25 Aug	0.288	1.06	1.05	-0.0041
26 Sep	0.274	1.32	1.30	-0.0066
30 Sep	0.274	1.55	1.36	-0.0568
2 Oct	0.264	1.20	1.30	0.0348
31 Oct	0.264	1.82	2.02	0.0453
1 Nov	0.268	1.40	1.89	0.1303
1 Nov	0.268	1.45	1.89	0.1151
19 Nov	0.268	2.35	2.50	0.0269
20 Nov	0.268	2.24	2.51	0.0494
3 Dec	0.278	3.00	2.99	-0.0015
4 Dec	0.278	2.60	3.01	0.0636
12 Dec	0.278	3.40	3.19	-0.0277
22 Dec	0.278	4.05	3.28	-0.0916
Sample arithmetic mean		2.13	2.07	-0.0032
Sample standard deviation (*S*)		1.04	0.84	0.0707

[1] Monthly mean O_3 from Nimbus-7 TOMS.

Step 7. This step tests the sample standard deviation for exceedence of the critical value specified in the original validity criterion. The hypotheses are phrased as

$$H_0: \sigma = \sigma^* = 0.1168, \text{ vs. } H_a: \sigma < 0.1168$$

In this test, the null hypothesis can be rejected if the computed χ^2 statistic is less than the value of χ^2 for the (1-α) level of significance with *n*-1 (ν=19) degrees of freedom (7.633).

The sample statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(19)(0.0707)^2}{(0.1168)^2} = 6.96$$

and, as 6.96 < 7.633, it can be concluded with 99% confidence that the precision of the model is adequate.

Step 8: Test the validity of the model, given criteria for acceptable accuracy and precision. Because the sample mean of the error observations is less than zero, the appropriate null hypothesis is

$$H_0: \mu_D = (-D^*+\delta) \text{ vs. } H_a: \mu_D > (-D^*+\delta)$$

In this case, $D^* = 0.301$, and, selecting the appropriate point on the *t* distribution to detect a 1% (one-sided) error rate,

$$\delta = \frac{t_{0.01,19}s}{\sqrt{n}} = \frac{(2.539)(0.0707)}{4.47} = 0.0402$$

The (single-sided) critical value t_c for 19 degrees of freedom (ν) at the α=0.01 level of significance is 2.539. The null hypothesis can be rejected if the computed *t* statistic is >2.529.

Table 7. χ^2 test of normality of error measure *D*

Intervals of 25% on *N(D;-0.0032, 0.0707)*	Obs Counts	Expect Prop	Exp Counts
< -0.0509	5	0.25	5
-0.0509 to -0.0032	5	0.25	5
-0.0032 to +0.0445	5	0.25	5
> +0.0445	5	0.25	5

The computed value is

$$t = \frac{(\bar{D} - (-D^* + \delta))\sqrt{n}}{s} = \frac{(-0.0032 - (-0.301 + 0.0402))\sqrt{20}}{0.0707} = 16.3$$

As $16.3 > t_c$ (=2.539), the null hypothesis can be rejected, and the simulation model can, with 99% confidence, be accepted as satisfying the performance criterion: better than 99% of EXAMS'

estimates differ by no more than 2× from experimental observations of clear-sky photolysis of DMDE.

Descriptive Statistics and Predictive Uncertainty

The uncertainty attaching to any point estimate of near-surface photolytic half-lives can be developed directly from the error measure *D* in **Table 6**. *D* has a mean value of -0.0032 and standard deviation 0.0707, giving a 99% confidence interval of -0.0032±0.0452. This interval encompasses uncertainties in the model, uncertainties in the measurement of light absorption spectra and quantum yields, and uncertainties in the experimental procedure used to measure DMDE photolysis. Because all these contribute to the uncertainty in final exposure estimates, it is appropriate to combine them into a single variance estimate.

The initial question was phrased in terms of exposure concentrations rather than the estimates of photolytic half-life that were used to test an internal constituent hypothesis of the EXAMS model. These specific results can, however, be used to project the uncertainty in photolysis kinetics onto the model exposure estimates used for risk assessment. EXAMS produces, as part of its standard outputs, estimates of acute and chronic mean exposure concentrations. Granting the average difference between the experimental and computed half-lives as deriving from model bias (rather than experimental error), the statistical properties of *D* can be translated into variation in predicted exposure concentrations via transformation of the distribution of half-lives into photolytic decay rate constants, and integration of the resulting chemical decay curves. By positing an example (albeit unrealistically simplified) scenario with an available analytical solution, the case can also be used to "verify" *sensu* [37] or "bench-mark" *sensu* [20] the numerical algorithms used in EXAMS to compute the time course of pesticide dissipation and to arrive at ecotoxicological exposure estimates.

For purposes of concrete illustration, consider the example of a chemical with an observed ("true") half-life of exactly 2 days, giving a disappearance rate constant *k* of 0.3466 d^{-1}. For a simple dissipation of DMDE from clean, shallow water, the average value over *t* days is $(C_0-C_t)/(kt)$, where C_0 is the initial concentration and C_t the concentration at time *t*. Taking the initial concentration as 1 mg/L, the "true" concentration after four days would be 0.25 mg/L, and the average (the 96-hour "acute exposure" value) would be 0.541 mg/L. Similarly, at the expiration of 21 days the average concentration (the "chronic" value) would be 0.137 mg/L. Evaluation of the predictive uncertainty of the model, in terms of the original requirement that exposure estimates be within a factor of two of reality at least 99% of the time, can now be accomplished. Examination of **Table 8** shows that EXAMS' estimates of acute (4-day) and chronic (21-day, 60-day, 90-day, and annual) exposures are well within the factor-of-two criterion at the end points of the 99% confidence limit on *D* (expressed as half-lives), thus confirming that the original requirement stated for testing predictions from the model has been satisfied.

Given its restricted range of process chemistry, this example validation is an illustration of the fact that global validation can only be accomplished through a series of specific validations, each of which is critically dependent on some subset of the constituent hypotheses of the model. In this instance, however, it has been unambiguously demonstrated that the EXAMS code can compute clear-sky radiative transfer and near-surface photolysis of synthetic chemicals, even under severe conditions of parameter approximation. Although EXAMS has many additional capabilities, each deserving specific validation, every instance of successful (objective, risk-controlled) validation adds to the confidence that can be placed in this simulation model, via its incremental Bayesian reduction in global model user's risk. Specifically, an accumulation of tests formulated to demonstrate, with 99% certainty, that the model is producing estimates within a factor of two of observed values more than 95% of the time, could be used to develop and support a variance or expected method error or intrinsic uncertainty for model point estimates used in probabilistic risk assessment.

Predictive Validity of Exposure Models

The quality and reliability of a model becomes established through a variety of tests. Internal tests of the quality of model construction are a necessary, but not a sufficient, condition for establishing a model's value as a regulatory tool: testing the strength of the welds in an automobile frame does not eliminate the need for full-scale crash tests. The only appropriate tests of a model's reliability as a predictive tool for regulatory decision ("validity") are those conducted by comparison of full model capabilities with independent experimental and field observations of model output quantities. Even so, the conclusions drawn must be qualified by the observation that any given test can only exercise a subset of the constituent hypotheses and computational techniques of the model.

Objective validations can only be conducted when the criteria for validity are objectively specified. Because the social consequences of accepting false models (inadequate chemical safety regulations) are much more serious than the consequences of rejecting true models (continued research and validation studies), model validations should always be phrased to test the null hypothesis that "the model is invalid." As only a single experimental unit is available for each measurement, paired sample techniques must be used to create a comparison metric *D*. When the underlying distribution of *D* can be identified, the tools of parametric hypothesis testing and statistical decision theory become available to the analyst, as in this validation of the EXAMS radiative transfer and direct photolysis algorithms. These tools allow for preliminary experimental designs guaranteed to control user's and modeler's risks at acceptable levels. In addition, the usual tools of statistical inference (confidence limits, etc.) can be used to compare models, to compute the validity properties (bias and dispersion) exhibited by a simulation model when it is subjected to particular experimental conditions, and to contribute to estimates of prediction uncertainty. Even when non-parametric tests must be

Table 8. Prediction uncertainty for rapid photolytic dissipation of DMDE from surface water

	Lower 99%	"True" Value	"True" Value	Upper 99%
Half-Life (days)	1.789	2.000	2.000	2.203
Source	EXAMS	Analytical	EXAMS	EXAMS
Acute (96-hour) Exposure (mg/L)	0.515	0.541	0.546	0.574
Chronic (21-day) Exposure (mg/L)	0.124	0.137	0.139	0.152
Average 60-day Exposure (mg/L)	0.0436	0.0481	0.0486	0.0534
Average 90-day Exposure (mg/L)	0.0290	0.0321	0.0324	0.0356
Average Annual Exposure (mg/L)	0.00716	0.00791	0.00798	0.00878

employed, however, the adverse consequences of using false models mandate that model user's risks be assigned to Type I errors by proper formulation of the null hypothesis as "this model is invalid" until proven otherwise.

It should also be remarked that this analysis has been formulated from an implicit view that the task of the model is to provide an adequate mimic of experimental results. In fact, what is sought is concordance between two views, the experimental and the mathematical, of a single natural object, in this instance chemical transformation in the presence of sunlight. Discord could accrue from errors in the mathematics, errors in the conduct of the experiments and measurements, or from errors in underlying physical and chemical conceptualizations. The similarity between complex numerical models and scientific theory is apparent. The continual attention to and repair of flaws discovered in numerical models differs little in motivation or technique from the more general development process of scientific theory.

Neither concordance nor dissonance between model and data should be taken as definitive, however, for notorious instances of the failure of each can be found in the history of science [14]. Astronomers of the 16th century, in response to Copernicus' heliocentric theory, searched for apparent movement of the stars ("stellar parallax") and, finding none, rejected Copernicus' theory based on the failure of observation to confirm theory. Today we realize that contemporary telescopes were inadequate to the task, and the failure was one of observation rather than one of theory. The wiser course, when confronted with a failure of observation to conform to theory, is often to reserve judgement.

In the 19th century, Lord Kelvin, working from well-established physical principles, calculated the age of the Earth from the rate of cooling of an initially molten sphere and concluded that Lyell's geology and Darwin's evolution by natural selection were invalid because Earth's history contained insufficient time for Uniformitarian processes to have brought about the observed world. This view held sway for a generation, and it was not until the discovery of radiogenic heat that evolutionary biology and geology began to regain their lost ground. The wiser course, when confronted with a failure of theory to conform to observation, is often to reserve judgement.

Finally, it should also be observed that 16th-century astronomers had a competing theory of the heavens at hand in the Ptolemaic view of the universe. Had that theory been computerized, its proponents would have been well-justified in terming it a "validated" model, with myriad instances of concordance between model prediction and astronomical observation. The wiser course, when confronted with a concordance of observation and theory, is often to proceed with regulatory action while always remembering the limits of scientific knowledge and retaining a willingness to revisit regulatory decisions in response to new discoveries and an evolving paradigm of risk assessment.

Current Model Validation Status

AgDisp/AgDrift

A systematic evaluation of the AgDisp algorithms, which simulate off-site drift and deposition of aerially applied pesticides, contained in the AgDRIFT® model [66] was performed by comparing model simulations to field trial data collected by the Spray Drift Task Force [67]. Field trial data used for model evaluation included 161 separate trials of typical agriculture aerial applications under a wide range of application and meteorological conditions. Input for model simulations included information on the aircraft and spray equipment, spray material, meteorology, and site geometry. The model input data sets were generated independently of the field deposition results – i.e., model inputs were in no way altered or selected to improve the fit of model output to field results. AgDisp shows a response similar to that of the field observations for many application variables (e.g., droplet size, application height, wind speed). AgDisp is, however, sensitive to evaporative effects, and modeled deposition in the far field responds to wet bulb depression although the field observations did not. The model tended to over-predict deposition rates relative to the field data

for far-field distances, particularly under evaporative conditions. AgDisp was in good agreement with field results for estimating near-field buffer zones needed to manage human, crop, livestock, and ecological exposure.

Pesticide Root Zone Model (PRZM)

The FIFRA Environmental Model Validation Task Force (FEMVTF), a collaborative effort of scientists from the crop protection industry and the U.S. Environmental Protection Agency, compared the results of PRZM predictions with measured data collected in 18 different leaching and runoff field studies as part of a process to improve confidence in the results of regulatory modeling; the following discussion is drawn from their report [68].

In its initial phase, the Model Validation Project reviewed existing, published studies on the validation of PRZM. The primary purpose of this literature review was to assess the quality and quantity of existing information to determine whether additional model validation studies were needed. A second purpose of the literature review was to collect information that would be useful in planning future model validation studies. The report [68] summarizes both aspects of this literature review and presents the reasons why the FIFRA Exposure Modeling Work Group concluded that more validation research would be useful in improving confidence in models used in regulatory assessments.

The literature search identified 35 articles involving the calibration/validation of model simulations with measured data. These studies included data from seven countries on three continents and a number of different compounds. Due to the varied nature of the papers and the lack of details for both model predictions and measured results, a detailed comparison of model predictions to observed data proved impossible. The majority of the papers indicated good agreement between model predictions and measurements, or that the models generally predicted more movement than actually occurred. These results, given the wide range of conditions reported in the papers, were taken by FEMVTF to lend general support to the use of PRZM in the regulatory process, especially for predicting leaching. Following review of this literature, the FIFRA Exposure Modeling Work Group decided that additional comparisons of field data and model predictions would be useful to supplement existing studies in helping improve confidence in the regulatory use of environmental models for predicting leaching and runoff. The following observations contributed to this decision:

· None of the published studies used the current version (3) of the model (this is especially important in that PRZM runoff routines have significantly evolved in version 3).

· Very few of the studies focused on runoff losses (most studies focused on the mobility of crop protection products in the soil profile).

· The number of studies having quantitative validation results is minimal. Since few of the published studies consider model validation the primary purpose of the field experiments, often data sets were not as extensive as would be desirable for model validation.

· Modelers were aware of field results in most of the studies (although in some of the studies where the field results were known, modelers claimed to make no adjustments to the input parameters). Therefore, in these studies the comparisons of model predictions and experimental measurements could be considered calibration – in model validation the modeler should have no knowledge of the field results to prevent biasing the selection of input parameters.

The Task Force report concludes that PRZM provides a reasonable estimate of chemical runoff at the edge of a field. Simulations based on the best choices for input parameters (no conservatism built into input parameters) were generally within an order of magnitude of measured data, with better agreement observed both for larger events and for cumulative values over the study period. When the model input parameters were calibrated to improve the hydrology, the fit between predicted values and observed data improved (results usually within a factor of three). When conservatism was deliberately introduced into the input pesticide parameters, substantial over-prediction of runoff loses occurred.

Simulations with PRZM obtained reasonable estimates of leaching in homogeneous soils where preferential flow was not significant. PRZM usually did a good job of predicting movement of bromide in soil (soil pore water concentrations were generally within a factor of two of predicted values). For simulations based on the best choices for input parameters (no conservatism built into input parameters), predictions of soil pore water concentrations for pesticides were usually within a factor of three. This was about a factor of two closer than when conservative assumptions were used to define input pesticide parameters. When the model input parameters were calibrated to improve the hydrology, predicted pesticide concentrations were usually within a factor of two of measured concentrations. Because of the sensitivity of leaching to degradation rate, the best predictions were obtained with pesticides with relatively slow degradation rates.

Differences in initial work conducted by different analysts demonstrated the importance of having a "standard operating procedure" to define the selection of *all* model input parameters. The FEMVTF concluded that the most satisfactory way to implement regulatory modeling is through the development of a "shell" that could provide all input parameters related to the scenario, with the user providing only the parameters related to the specific pesticide being assessed.

Exposure Analysis Modeling System (EXAMS)

Validation studies of EXAMS have been conducted, *inter alia*, in the Monongahela River, USA [69], an outdoor pond in Germany [70], a bay (Norrsundet Bay) on the east coast of Sweden [71, 72], Japanese rice paddies [73], and rivers in the UK [74] and in South Dakota [75].

In the Monongahela River study [69], model predictions of phenol concentration downstream from a steel mill effluent were compared to ambient data. Agreement between observed levels of phenol and model predictions was best when the concentration of oxidizing species was treated as a reach-specific calibration parameter, although satisfactory agreement was evident using a single value of 10^{-8} M.

The dyestuff Disperse Yellow (DY 42, C.I. No. 10338) was introduced into an outdoor pond [70]. The model was judged to show "good agreement" with the measured behavior of the dye, but the published concentration time-series clearly indicates that the exchange rate used in the authors' pond scenario underestimated the velocity of capture of the dye by benthic sediments.

Norrsundet Bay is heavily polluted with kraft mill effluent [71, 72]. The model scenario was calibrated using data on chloroform in wastewater and seawater, and then tested on four other pollutants present in the wastewater (2,4,6-trichlorophenol, 3,4,5-trichloroguaiacol, tetra-chloroguaiacol, and tetrachlorocatechol). The results were judged "satisfactory" for three of the test compounds; failures with tetrachlorocatechol were suspected to arise from this compound's high affinity for suspended sediment.

In studies of the dissipation of sulfonylurea herbicides in rice paddy [73], model scenarios were calibrated against paddy water dissipation data from two 1-m^2 simulated paddies to obtain values for the benthic exchange coefficient and oxidative radical concentrations. The EXAMS scenario then successfully predicted the partitioning and degradation that led to half-lives in paddy water of 3 – 4 days observed in field studies conducted in Japan.

River studies in the UK [74] compared model predictions to downstream losses of styrene, xylenes, dichloro-benzenes, and 4-phenyldodecane in treatment plant effluent. Quantitative predictions compared well with observed values for those compounds for which reliable environmental rate data were available. Rapid losses of 4-phenyldodecane were observed in the river, but there were not sufficient data on the postulated loss mechanism (indirect photolysis) to derive input parameters for the model. Fate and transport of an anionic surfactant were studied in Rapid Creek, South Dakota [75]. Laboratory limnetic and benthic biolysis constants were used, and benthic exchange was calibrated to field data. EXAMS predictions agreed with observed water column and sediment concentrations to within factors of ±2 and ±4 respectively.

Validation studies of 1,4-dichlorobenzene in Lake Zurich, Switzerland [32] showed excellent agreement (differences <10%) between values measured in the Lake and EXAMS predictions; this study is included as an example application in the EXAMS user's guide [4]. Volatilization of diazinon, parathion, methyl parathion, and malathion from water, wet soil, and a water-soil mixture were studied in a simple environmental chamber and the results compared to EXAMS estimates of volatilization rates [76]. Experimental and predicted daily fractions volatilized agreed within a factor of three for diazinon, methyl parathion, and malathion, and within a factor of five for parathion, despite the fact that EXAMS was not designed for use with wet soil systems.

Bioaccumulation and Aquatic System Simulator (BASS)

BASS' bioconcentration and bioaccumulation algorithms have been validated by comparing its predicted uptake and elimination rates to values published in the peer-reviewed literature [77, 78]. These comparisons encompass both a wide variety of fish species, including Atlantic salmon (*Salmo salar*), brook trout (*Salvelinus fontinalis*), brown trout (*Salmo trutta*), carp (*Cyprinus carpio*), fathead minnow (*Pimephales promelas*), flagfish (*Jordanella floridae*), goldfish (*Carassius auratus*), golden orfe (*Leuciscus idus*), guppy (*Poecilia reticulata*), killifish (*Oryzias latipes*), mosquitofish (*Gambusia affinis*), and rainbow trout (*Oncorhynchus mykiss*), as well as a wide variety of chemical classes (brominated benzenes, brominated toluenes, chlorinated anisoles, chlorinated benzenes, chlorinated toluenes, organo-phosphorus pesticides, polybrominated biphenyls (PBBs), poly-chlorinated aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), polychlorinated dibenzofurans, polychlorinated dibenzodioxins, polychlorinated diphenyl ethers, polychlorinated insecticides, etc.). Reduced major axis regression analysis of observed vs. predicted exchange rates demonstrated excellent agreement between predicted rates of gill uptake and elimination vs. published values.

For organic chemicals FGETS/BASS bioaccumulation algorithms have also been validated by simulations of mixtures of PCBs in Lake Ontario salmonids and laboratory studies [79]. In these studies FGETS/BASS simulations of PCBs in Lake Ontario salmonids agreed well with observed data, and FGETS/BASS correctly simulated the relative contribution of gill and dietary routes of exposure for such hydrophobic chemicals as polychlorinated dibenzodioxins. For sulfhydryl-binding metals, BASS' bioconcentration algorithms have been validated by simulations of methyl mercury bioaccumulation in Florida Everglades fish communities [80]. As with the Lake Ontario PCB simulations, BASS methyl mercury simulations of fish communities in the Florida Everglades agreed well with observed data. Validation studies of BASS's bioenergetic growth algorithms are also available [79, 80].

DataBase Documentation

Accurate Monte Carlo simulation requires careful attention to covariance among input parameters driving the models. The assumption of independence among variables that are in fact correlated has on some occasions been found to artificially deflate, and on others to inflate, the variability observed in model outputs. "Trace matching" is the term used for parameter inputs that are based on observational data absent any attempt to fit distribution functions to the data. This technique is particularly attractive for geographic and climatological input datasets, for reasons both of preserving an accurate representation of extreme events and for robust preservation of covariances among variables. Several datasets for driving simulation models are under development; this section documents their sources and some aspects of initial database design.

Agricultural Geography

The geographic base for developing pesticide exposure datasets is the intersection of MLRA (Major Land Resource Areas) and state boundaries (**Figure 12**). The Soil Conservation Service (SCS), in its Agriculture Handbook 296 (AH-296), published a map and accompanying text description delineating 20 Land Resource Regions (LRR) of the coterminous United States [81]. Each LRR is subdivided into Major Land Resource Areas (MLRA); these are the fundamental geography of physiographic units used in this project. The LRR and MLRA were updated in 1984 through release of a Digital Line Graph (DLG) of the MLRA by the National Cartographic Center (Fort Worth, Texas) of the Soil Conservation Service. This update included major and minor boundary changes, and the elimination of duplicate MLRA in LRR (e.g., 048A occurs only in LRR D and was removed from LRR E).

AH-296 assembled available information about the land as a resource for farming, ranching, forestry, engineering, recreation, and other uses. The 1981 release of AH-296 improved upon its predecessor (AH-296 of 1965) in its refined cartography, identification of the soils present in each MLRA, description of the potential natural vegetation of each, and the addition of Alaska, Hawaii, and Puerto Rico to the inventory. AH-296 was designed as a resource for making decisions in state-wide, interstate, regional, and national agricultural planning, as a base for natural resource inventories, as a framework for extrapolating results within physiographic units, and for organizing resource conservation programs.

The land resource categories used at state and national levels are, in increasing geographical extent, land resource *units*, land resource *areas*, and land resource *regions*. Land resource *units* are geographic areas, usually a few thousand hectares in extent, that are characterized by a particular pattern of soils, climate, water resources, and land use. A *unit* can be a single continuous area or several separate nearby areas. Land resource units (LRU) are the basic units from which major land resource *areas* are determined. LRU are also the basic units for state land resource maps. They are usually co-extensive with state general soil map units, although some general soil map units are subdivided to produce LRU because of significant geographic distinctions in climate, water resources, or land use. Major Land Resource *Areas* (MLRA) are geographically associated land resource units (LRU). In AH-296, MLRA are further grouped into Land Resource *Regions* (LRR) that are designated by capital letters and identified with a descriptive geographical name (see **Figure 12**). For example, the descriptive name for Land Resource Region *S* is the "Northern Atlantic Slope Diversified Farming Region."

The MLRA subdivisions of the LRR range in number from a minimum of 2 (Alluvium; Silty Uplands) MLRA divisions in Region *O* (Mississippi Delta Cotton and Feed Grains Region), to a high of 23 MLRA in Region *D* (Western Range and Irrigated Region). In land area, the 181 MLRA range from a low of 2,476 km^2 (C-16, California Delta sector of Region C – California Subtropical Fruit, Truck, and Specialty Crop Region), to the 294,252 km^2 of the Northern Rocky Mountains (E-43, in the Rocky Mountain Range and Forest Region), with a median size of 30,080 km^2. Outlines of the MLRA are indicated in **Figure 12**.

Physiography of LRR and MLRA

AH-296 briefly describes the dominant physical characteristics of the Land Resource Regions and the Major Land Resource Areas under the headings of land use, elevation and topography, climate, water, soils, and potential natural vegetation:

Land use – The extent of the land used for cropland, pasture, range, forests, industrial and urban development, and other special purposes is indicated. A list of the principal crops grown and the type of farming practiced is included.

Elevation and topography – a range in elevation above sea level and any significant exceptions is provided for the MLRA as a whole. The topography of the area is described.

Climate – AH-296 gives a range of the annual precipitation for

the driest parts of the area to the wettest and the seasonal distribution of precipitation, plus a range of the average annual temperature and the average freeze-free period characteristic of different parts of the MLRA.

Water – Information is provided on surface stream-flow and ground water, and the source of water for municipal use and irrigation. MLRAs dependent on other areas for water supply and those supplying water to other areas are identified.

Soils – The dominant soils are identified according to the principal suborders, great groups, and representative soil series.

Potential natural vegetation – the plant species that the MLRA can support are identified by their common names.

Meteorology: SAMSON/HUSWO

Although "weather generator" software is available and is regularly used for water resource and climate studies, it has been observed that the weather sequences thus generated are weak in their ability to capture the extreme events that are usually of greatest importance in risk assessments. In a study [82] of the USCLIMATE [83, 84] and CLIGEN [85] models, the authors remark that "Annual and monthly precipitation statistics (means, standard deviations, and extremes) were adequately replicated by both models, but daily amounts, particularly typical extreme amounts in any given year, were not entirely satisfactorily modeled by either model. USCLIMATE consistently underestimated extreme daily amounts, by as much as 50%.". In a study [86] of WGEN [87] (itself an element of USCLIMATE) and LARS-WG [88] at 18 climatologically diverse sites in the USA, Europe, and Asia, the authors conclude that the gamma distribution used in WGEN "probably tends to overestimate the probability of larger values" of rainfall. This result, although opposite in tendency to that of [82], is no less undesirable. Both models had a lower inter-annual variance in monthly mean precipitation than that in the observed data, and neither generator "performed uniformly well in simulating the daily variances of the climate variables."

Issues of accurately preserving the covariance among parameters can be completely by-passed by using observed synoptic data, and all danger of generating impossible input scenarios (e.g., days of heavy rainfall coupled with maximum drying potential) can be avoided, given adequate quality assurance of the input datasets. Weather generator software can be problematic in this regard; for example, CLIGEN generates temperature, solar radiation, and precipitation independently of one another, so the covariance structure of daily sequences is clearly not preserved in the model outputs. Semenov et al. [86] concluded that failures to represent variance in LARS-WG and WGEN were "likely to be due to the observed data containing many periods in which successive values are highly correlated..."

Data from SAMSON (Solar and Meteorological Surface Observation Network) is available as a three-volume CD-ROM disk set that contains observational and modeled meteorological and solar radiation data for the period 1961-1990. An additional CD-ROM (HUSWO, Hourly United States Weather Observations) extends the data set to 1995. Combined data are available for 234 National Weather Service stations in the United States, Guam and Puerto Rico (**Figure 13**). Appendix A lists the available stations by their standard WBAN (Weather Bureau Army Navy) number, with station location (City, State), geographic (latitude and longitude) coordinates, and station elevation (m).

The hourly SAMSON solar elements are: extraterrestrial horizontal and extraterrestrial direct normal radiation; and global, diffuse, and direct normal radiation. Meteorological elements are: total and opaque sky cover, temperature and dew point, relative humidity, station pressure, wind direction and speed, visibility, ceiling height, present weather, precipitable water, aerosol optical depth, snow depth, days since last snowfall, and hourly precipitation. An additional five years of data (1991-1995) were acquired on CD from NCDC (National Climatic Data Center) as the HUSWO (Hourly United States Weather Observations) product. HUSWO is an update to the SAMSON files. Weather elements in the files include total and opaque sky cover; temperature and dew point; relative humidity; station pressure; wind direction and speed; visibility; ceiling height; present weather; ASOS cloud layer data; snow depth; and hourly precipitation for most stations. Stations for which hourly precipitation is unavailable are indicated in the station list of Appendix A.

Soils and Land Use

The National Resources Inventory (NRI) is a statistically-based inventory of land cover and use, soil erosion, prime farmland, wetlands, and other natural resource characteristics on non-Federal rural land in the United States. Inventories are conducted at 5-year intervals by the U.S. Department of Agriculture's Natural Resources Conservation Service (NRCS, formerly the Soil Conservation Service (SCS)), to determine the condition and trends in the use of soil, water, and related resources nationwide and statewide.

The 1992 NRI covered 170 data elements at some 800,000 sample points on the 1.5 billion acres of non-Federal land in the USA – some 75 percent of the Nation's land area. At each sample point, information is available for three years-1982, 1987, and 1992. Data is currently being re-summarized for 1997. Originated as a means of getting accurate natural resource information to USDA policymakers, the NRI has become useful to a variety of users. The NRI contains three codes identifying the geographic location of each point by its Major Land Resource Area (MLRA), Hydrologic Unit Code (HUC), and County (represented by five-digit codes). The MLRAs are geographically-associated land resource units, which in turn are geographic areas, usually several thousand acres in extent, characterized by a particular pattern of soils, climate, water resources, and land use. Hydrologic Unit Codes (HUC) consist of eight digits denoting major stream drainage basins as defined and digitized by the U.S. Geological Survey. County five-digit codes are

standard FIPS (Federal Information Processing Standards) identifiers in which the first two digits identify the State and the remaining three the individual Counties. With these geographies combined in a GIS and with the Federal lands masked out, the individual regions represent the smallest spatial feature that can be used to locate NRI samples. Privacy concerns preclude detailed analysis at this finest level of spatial resolution, however. NRI data are statistically reliable for national, regional, state, and sub-state analysis; for this project the state-level sections of MLRAs have been chosen for summary. The NRI was scientifically designed and conducted, and is based on recognized statistical sampling methods. The data are used in national, state, and local planning, university research, and private sector analysis. NRI data help shape major environmental and land-use decisions, and hold considerable potential for contributing to analysis of potential pesticide off-site migration, fate and effects.

National Resource Inventory Data Characteristics

Data collected in the 1982, 1987, 1992 and 1997 NRI provide a basis for analysis of 5-year and 10-year trends in resource conditions. Many data items in the 1997 NRI are consistent with previous inventories. In addition, the NRI is linked to the Natural Resources Conservation Service's extensive Soil Interpretations Records to provide additional soils information suitable for the PRZM model.

Data elements consistent within the NRI database are:

- Farmstead, urban, and built-up areas
- Farmstead and field windbreaks
- Streams less than 1/8 mile wide and water bodies less than 40 acres
- Type of land ownership
- Soils information – soil classification, soil properties, and soil interpretations such as prime farmland
- Land cover/use – cropland, pasture land, rangeland, forest land, barren land, rural land, urban and built-up areas

The *cropland* land cover/use category includes areas used for the production of adapted crops for harvest, including row crops, small grain crops, hay crops, nursery crops, orchard crops, and other specialty crops. *Cultivated cropland* includes land identified as being in row or close-grown crops, summer fallow, aquaculture in crop rotation, hayland or pastureland in a rotation with row or close grown crops, or horticulture that is double cropped. Also included is "cropland not planted" because of weather conditions, or because the land is in a USDA set-aside or similar short-term program, or because of other short-term circumstances. *Non-cultivated cropland* includes land that is in a permanent hayland or horticultural crop cover; hayland that is managed for the production of forage crops (grasses, legumes) that are machine harvested; horticultural cropland that is used for growing fruit, nut, berry, vineyard, and other bush fruit, and similar crops. Cropland information represented includes:

- Cropping history
- Irrigation-type and source of water
- Erosion data-wind and water
- Wetlands-classification of wetlands and deepwater habitats in the U.S. (not in 1987)
- Conservation practices and treatment needed
- Potential conversion to cropland
- Rangeland condition, apparent trend of condition

New data elements added for the 1992 NRI included:

- Streams greater than 1/8 mile wide and water bodies by kind and size greater than 40 acres
- Conservation Reserve Program land under contract
- Type of earth cover – crop, tree, shrub, grass-herbaceous, barren, artificial, water
- Forest type group
- Primary and secondary use of land and water
- Wildlife habitat diversity
- Irrigation water delivery system
- Food Security Act (FSA) wetland classification
- For rangeland areas – range site name and number, woody canopy, noxious weeds
- Concentrated flow, gully, and streambank erosion
- Conservation treatment needed
- Type of conservation tillage

Stratospheric Ozone from the TOMS

Photochemical transformation of pesticides is driven by sunlight reaching the surface of the Earth. Higher-energy wavelengths in the ultraviolet portion of the solar spectrum are in most cases the most potent for effecting these transformations. Exposure models that estimate ground-level solar spectral intensity require access to stratospheric ozone data as an input to calculations of losses in incoming radiation during passage through the atmosphere. To meet this requirement, a world-wide database was developed from the latest release (1996, using Version 7 of the data reduction algorithm) of ozone data from the TOMS (Total Ozone Mapping Spectrometer) instrument flown on the Nimbus-7 spacecraft [89]. The dataset was derived from a 2 CD-ROM set containing data covering the entire Nimbus-7 TOMS lifetime (November 1, 1978 through May 6, 1993), given as monthly averages.

The Ozone Measurement

The Nimbus-7 spacecraft was in a south-to-north, sun-synchronous polar orbit so that it was always close to local noon/midnight below the spacecraft. Thus, ozone measurements were taken for the entire world every 24 hours. TOMS directly measured the ultraviolet sunlight scattered by the Earth's atmosphere. This NASA-developed instrument measured ozone indirectly by mapping ultraviolet light emitted by the Sun to that scattered from the Earth's atmosphere back to the satellite. Total column ozone was inferred from the differential absorption of scattered sunlight in the ultraviolet range. Ozone was calculated by taking the ratio of two wavelengths (312 nm and 331 nm, for

example), where one wavelength is strongly absorbed by ozone while the other is absorbed only weakly. The instrument had a 50 kilometer square field of view at the sub-satellite point. TOMS collected 35 measurements every 8 seconds as it scanned right to left, producing approximately 200,000 ozone measurements daily. These individual measurements varied typically between 100 and 650 Dobson Units (DU) and averaged about 300 DU. This is equivalent to an 0.3 cm (about a 10th of an inch) thick layer of pure ozone gas at NTP (Normal Temperature and Pressure).

The Data Files

Gridded Monthly Average

For each month, the individual TOMS measurements were averaged into grid cells covering 1 degree of latitude by 1.25 degrees of longitude. The 180x288 ASCII data array contains data from 90S to 90N, from 180W to 180E. Each ozone value is a 3 digit integer. For each grid cell, at least 20 days of data in any given month and year were required to be good, for the monthly average to have been computed. For pesticide exposure modeling, these files were averaged to provide grand mean monthly values for the period of record for each grid cell.

Zonal Means

Monthly zonal means were available from the file \zonalavg\zonalmon.n7t on the 2nd CD. The averages are for 5 degree latitude zones, area-weighted. At least 75% of possible data in a given zone was required to be present for the mean to be calculated. In 1978 and 1979 there were missing days when the TOMS instrument was turned off to conserve power. In the later years there are at least some data every day. The units of measurement for the zonal means are Dobson Units. EXAMS' ozone module defaults to zonal means when the data file cannot be found.

Problems with the Data

Polar Night TOMS measured ozone using scattered sunlight; it is not possible to measure ozone when there is no sun (in the polar regions in winter). Consequently, for example, the Antarctic polar regions for August and September always have areas of missing data due to polar night. These gaps were filled by the expedient of averaging the monthly zonal means across all available years, interpolating from polar dusk to polar dawn during periods of continuous darkness, and then substituting these values for zeros remaining in the monthly gridded dataset after incorporating all available monthly gridded data.

Missing Data During 1978 and 1979 the TOMS instrument was turned off periodically to conserve power, including a 5-day period (6/14- 6/18) in June 1979. On many days, data were lost due to missing orbits or other problems. The sample size among grid cells is thus not identical. The variance (2 S.E.) in the ozone data over the 14-year lifetime of the instrument is, however, only 1.5%.

High Terrain The ozone reported is total column ozone to the ground. Over high mountains (the Himalayas, the Andes) low ozone will be noticed relative to surrounding low terrain. This is not an error.

Pesticide Usage

Current pesticide use patterns and rates can be of value for evaluating exposure of entire classes of compounds (e.g., the organophosphorus insecticides) or the actual usage pattern of single registered compounds.

Both EPA [90, 91] and USDA accumulate data on the sale and use of pesticides. USDA pesticide use surveys include eight benchmark years (1964, 1966, 1971, 1976, 1982, 1990, 1991, and 1992) [92]. Consistent information over time is, however, only available for eleven crops: corn, cotton, soybeans, wheat, rice, grain sorghum, peanuts, fall potatoes, other vegetables, citrus, and apples. Under the sponsorship of EPA, USDA, and the Water Resources Division of USGS, the National Center for Food and Agricultural Policy (NCFAP) assembled a comprehensive database of pesticide use in American agriculture [93]. The NCFAP database is not specific to any particular year; it is a summary compilation of studies conducted by public agencies over the four-year period 1990-1993, including

- National Agricultural Statistics Service (NASS) surveys of pesticide use in field crops, vegetable crops, and fruit and nut crops,
- reports funded by the USDA Cooperative Extension Service (CES),
- pesticide benefit assessments from the USDA National Agricultural Pesticide Impact Assessment Program (NAPIAP), and
- State of California compilations of farmers' pesticide use records, supplemented by
- NCFAP surveys of Extension Service specialists, and,
- where necessary, imputations developed from the assumption that neighboring States' pesticide use profiles are similar.

The 15,740 individual use records in the database – covering 200 active ingredients (a.i) and 87 crops – are State-level point estimates focused on two use coefficients: (1) the percent of a crop's acreage in a State treated with an individual a.i., and (2) the average annual application rate of the active ingredient per treated acre.

Because these data represent the average application and treatment rates by State, they do not yield precise estimates of use at the sub-State level. The State use coefficients represent an average for the entire State and consequently do not reflect the local variability of cropping and management practices. This is an irreducible uncertainty not readily amenable to quantification. The reliability of these (State-level) estimates can, however, be evaluated from NASS (National Agricultural Statistics Service) assessments of the coefficient of variation (c.v.) or percentage

relative standard error (% rse) of data in the NASS chemical usage reports used in building the NCFAP database [94-96]. The variability due to sampling error is calculated for all chemical and acreage variables in the NASS surveys, and expressed as a percentage of the estimate. **Table 9** shows the entire range of sampling variability for percent of acres treated and application rate for each crop class (field crops, fruits, vegetables) across all crops and States surveyed. The particular value to be used should be selected based upon the number of reports used to develop an estimate for a particular crop in a particular State; for aggregated totals (e.g., all field crops within a multi-State region, combined usage of organo-phosphorus insecticides within an entire State, etc.), the combined sample size should be used to select the appropriate % rse.

These data can be used to calculate approximate confidence limits, and to evaluate the significance of inter-annual variability in the source data underlying the NCFAP database. In general the % rse can be interpreted by imagining that the surveys are repeated many times using the same sample size: in two out of three cases, the outcome would not differ from the database value by more than the stated sampling variability. Approximate confidence bands can be calculated by applying values from Table 2 to NCFAP data elements. For example, if a tabulated value gives 20% of a field crop treated with a specific pesticide, the (66%) confidence band for a State with few reports would be 20±(20×0.35) or 20±7 percent of the crop acreage. For a State with a large sample size, the confidence interval would be 20±(20×0.10) or 20±2 percent of the crop acreage. For comparison of values, an overlap of confidence bands at twice the % rse (i.e., 2 standard errors) indicates that the estimates have only a 1 in 20 chance of being genuinely different.

Table 9. Reliability statement (sampling variability expressed as percentage relative standard error (rse) of the estimate) of 1992 pesticide data (Fruits survey conducted in 1993 crop year) in NASS agricultural chemical usage reports. The rse to be applied to a specific datum depends on the size of the sample used to develop the item; the table entries indicate the *range* of rse encountered in the data sets

Tabulated % acres treated	Field Crops (1992 Crop Year)		Fruits (1993 Crop Year)		Vegetables (1992 Crop Year)	
	Acres Treated	Appl. Rate	Acres Treated	Appl. Rate	Acres Treated	Appl. Rate
< 10	40-100	1-60	25-90	1-30	35-85	1-10
10 - 24	10-35	5-35	15-65	1-20	20-70	1-10
25 - 49	5-15	1-30	10-35	1-20	10-40	1-10
50-75	5-15	5-25	5-20	1-15	5-20	1-10
> 75	1-5	1-10	1-10	1-5	1-5	1-10

State Parts of MLRA

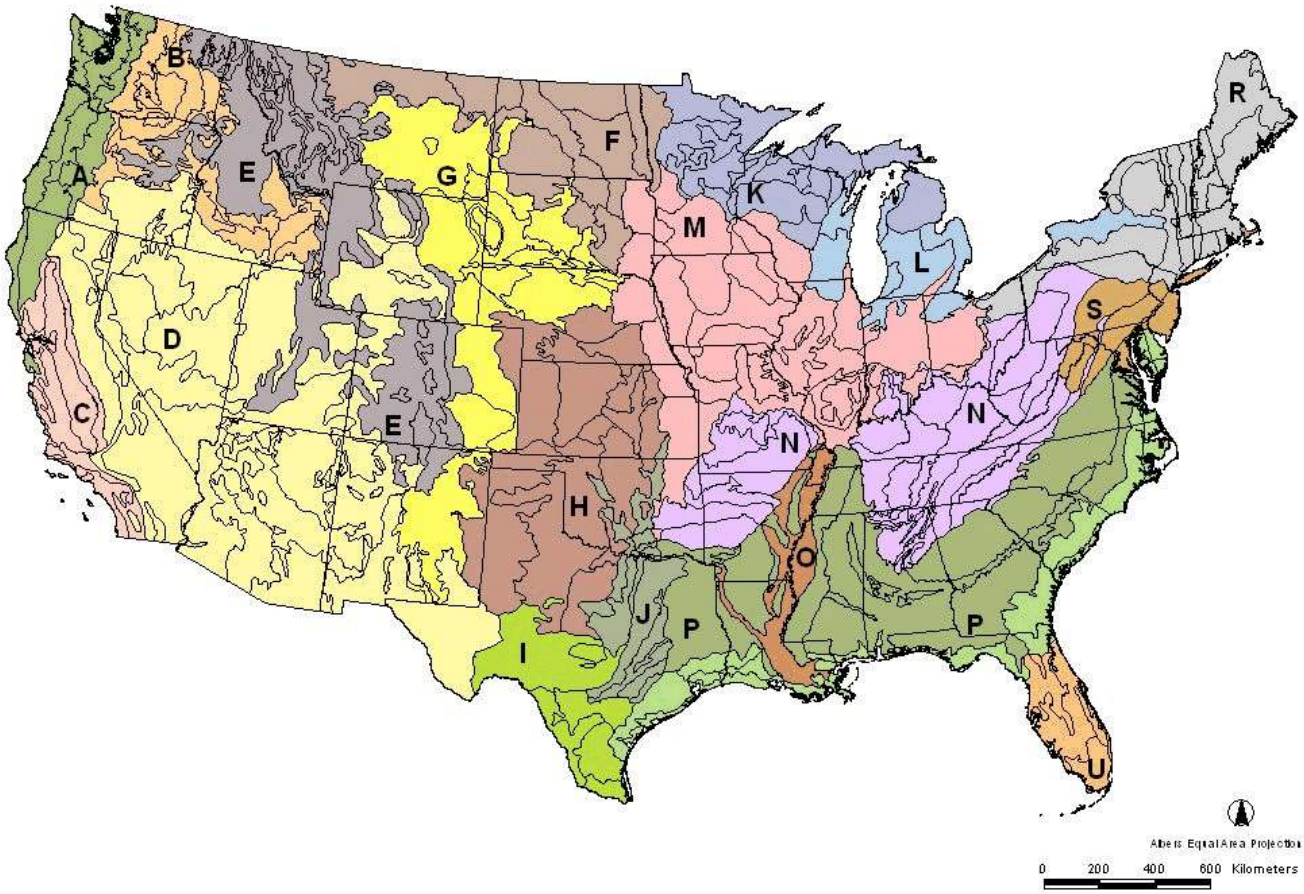

Figure 12. Land Resource Regions and Major Land Resource Areas with State Boundaries.

LAND RESOURCE REGIONS

- A Northwestern Forest, Forage, and Specialty Crop Region
- B Northwestern Wheat and Range Region
- C California Subtropical Fruit, Truck, and Specialty Crop Region
- D Western Range and Irrigated Region
- E Rocky Mountain Range and Forest Region
- F Northern Great Plains Spring Wheat Region
- G Western Great Plains Range and Irrigated Region
- H Central Great Plains Winter Wheat and Range Region
- I Southwest Plateaus and Plains Range and Cotton Region
- J Southwestern Prairies Cotton and Forage Region
- K Northern Lake States Forest and Forage Region
- L Lake States Fruit, Truck, and Dairy Region
- M Central Feed Grains and Livestock Region
- N East and Central Farming and Forest Region
- O Mississippi Delta Cotton and Feed Grains Region
- P South Atlantic and Gulf Slope Cash Crops, Forest, and Livestock Region
- R Northeastern Forage and Forest Region
- S Northern Atlantic Slope Diversified Farming Region
- T Atlantic and Gulf Coast Lowland Forest and Crop Region
- U Florida Subtropical Fruit, Truck Crop, and Range Region

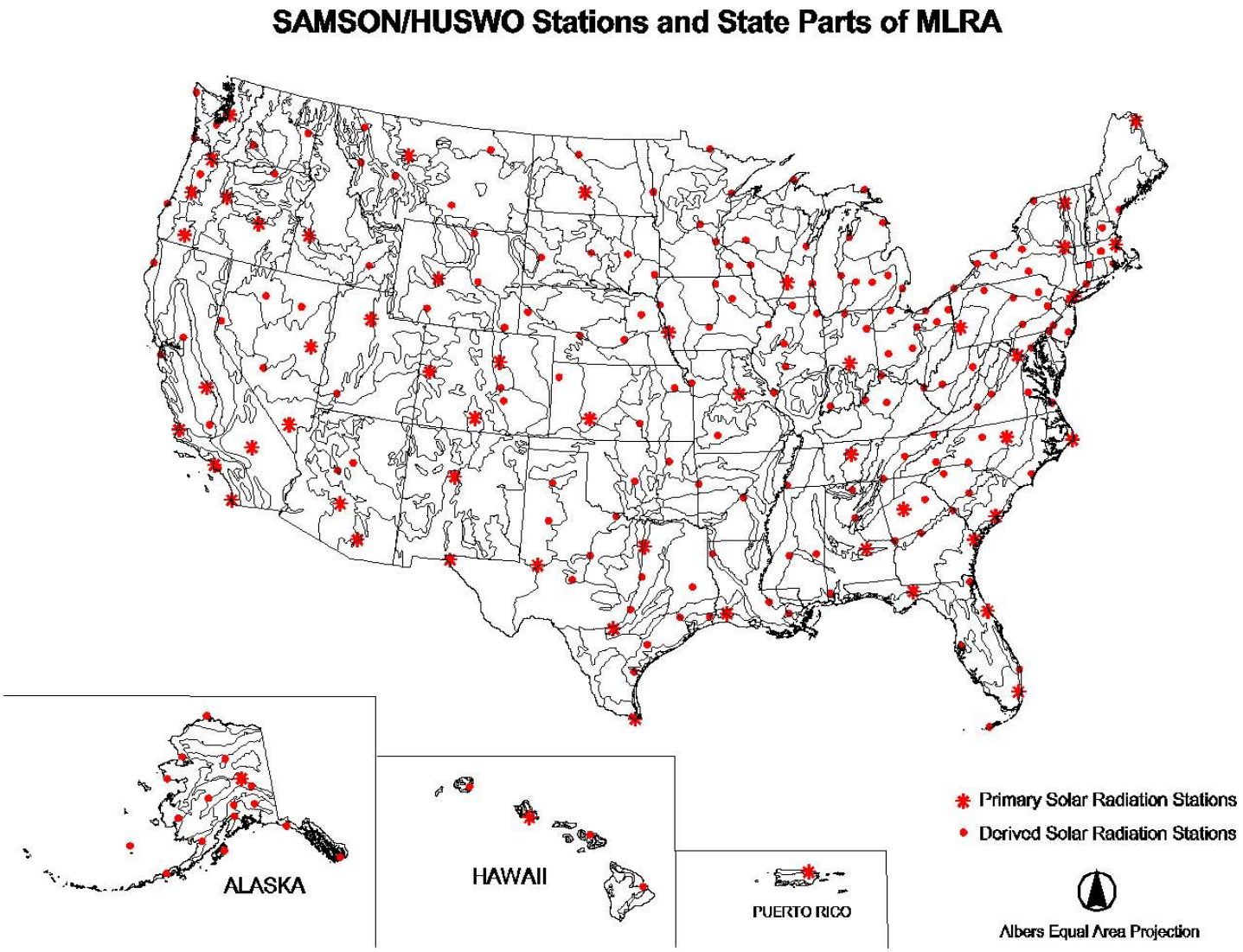

Figure 13. Location of SAMSON/HUSWO Stations for Integrated Climatological Database.

Appendix: SAMSON/HUSWO Stations

Weather Bureau Army Navy (WBAN) number, station location (City and State), geographic coordinates (latitude and longitude) and elevation (m M.S.L.) of the 234 stations available for coordinated climatological dataset for AgDisp, PRZM, and EXAMS (see **Figure 13**). Primary stations (those with measured solar radiation data for at least one year) are in bold type; stations lacking hourly precipitation data are italicized.

WBAN	Station	State	Lat.	Long.	Elev. (m)
03103	Flagstaff	AZ	35.1	-111.7	2135
03812	Asheville	NC	35.4	-82.5	661
03813	Macon	GA	32.7	-83.7	110
03820	Augusta	GA	33.4	-82.0	45
03822	**Savannah**	GA	32.1	-81.2	16
03856	Huntsville	AL	34.7	-86.8	190
03860	Huntington	WV	38.4	-82.6	255
03870	Greenville	SC	34.9	-82.2	296
03927	**Fort Worth**	TX	32.8	-97.1	164
03928	Wichita	KS	37.7	-97.4	408
03937	**Lake Charles**	LA	30.1	-93.2	3
03940	Jackson	MS	32.3	-90.1	101
03945	**Columbia**	MO	38.8	-92.2	270
03947	Kansas City	MO	39.3	-94.7	315
04725	Binghamton	NY	42.2	-76.0	499
04751	Bradford	PA	41.8	-78.6	600
11641	**San Juan**	PR	18.4	-66.0	19
12834	**Daytona Beach**	FL	29.2	-81.1	12
12836	Key West	FL	24.6	-81.8	1
12839	**Miami**	FL	25.8	-80.3	2
12842	Tampa	FL	28.0	-82.5	3
12844	West Palm Beach	FL	26.7	-80.1	6
12912	Victoria	TX	28.9	-96.9	32
12916	New Orleans	LA	30.0	-90.3	3
12917	Port Arthur	TX	30.0	-94.0	7
12919	**Brownsville**	TX	25.9	-97.4	6
12921	**San Antonio**	TX	29.5	-98.5	242
12924	Corpus Christi	TX	27.8	-97.5	13
12960	Houston	TX	30.0	-95.4	33
13722	**Raleigh**	NC	35.9	-78.8	134
13723	Greensboro	NC	36.1	-80.0	270
13729	Elkins	WV	38.9	-79.9	594
13733	Lynchburg	VA	37.3	-79.2	279
13737	Norfolk	VA	36.9	-76.2	9
13739	Philadelphia	PA	39.9	-75.3	9
13740	Richmond	VA	37.5	-77.3	50
13741	Roanoke	VA	37.3	-80.0	358
13748	Wilmington	NC	34.3	-77.9	9
13781	Wilmington	DE	39.7	-75.6	24
13865	Meridian	MS	32.3	-88.8	94

WBAN	Station	State	Lat.	Long.	Elev. (m)
13866	Charleston	WV	38.4	-81.6	290
13873	Athens	GA	34.0	-83.3	244
13874	**Atlanta**	GA	33.7	-84.4	315
13876	Birmingham	AL	33.6	-86.8	192
13877	Bristol	TN	36.5	-82.4	459
13880	**Charleston**	SC	32.9	-80.0	12
13881	Charlotte	NC	35.2	-80.9	234
13882	Chattanooga	TN	35.0	-85.2	210
13883	Columbia	SC	34.0	-81.1	69
13889	Jacksonville	FL	30.5	-81.7	9
13891	Knoxville	TN	35.8	-84.0	299
13893	Memphis	TN	35.1	-90.0	87
13894	Mobile	AL	30.7	-88.3	67
13895	**Montgomery**	AL	32.3	-86.4	62
13897	**Nashville**	TN	36.1	-86.7	180
13957	Shreveport	LA	32.5	-93.8	79
13958	Austin	TX	30.3	-97.7	189
13959	Waco	TX	31.6	-97.2	155
13962	Abilene	TX	32.4	-99.7	534
13963	Little Rock	AR	34.7	-92.2	81
13964	Fort Smith	AR	35.3	-94.4	141
13966	Wichita Falls	TX	34.0	-98.5	314
13967	Oklahoma City	OK	35.4	-97.6	397
13968	Tulsa	OK	36.2	-95.9	206
13970	Baton Rouge	LA	30.5	-91.2	23
13985	**Dodge City**	KS	37.8	-100.0	787
13994	St. Louis	MO	38.8	-90.4	172
13995	Springfield	MO	37.2	-93.4	387
13996	Topeka	KS	39.1	-95.6	270
14607	**Caribou**	ME	46.9	-68.0	190
14733	Buffalo	NY	42.9	-78.7	215
14734	Newark	NJ	40.7	-74.2	9
14735	**Albany**	NY	42.8	-73.8	89
14737	Allentown	PA	40.7	-75.4	117
14739	**Boston**	MA	42.4	-71.0	5
14740	Hartford	CT	41.9	-72.7	55
14742	**Burlington**	VT	44.5	-73.2	104
14745	Concord	NH	43.2	-71.5	105
14751	Harrisburg	PA	40.2	-76.9	106
14764	Portland	ME	43.7	-70.3	19
14765	Providence	RI	41.7	-71.4	19
14768	Rochester	NY	43.1	-77.7	169
14771	Syracuse	NY	43.1	-76.1	124
14777	Wilkes-Barre	PA	41.3	-75.7	289
14778	*Williamsport*	PA	41.3	-77.1	243
14820	Cleveland	OH	41.4	-81.9	245
14821	Columbus	OH	40.0	-82.9	254
14826	Flint	MI	43.0	-83.7	233
14827	Fort Wayne	IN	41.0	-85.2	252
14836	Lansing	MI	42.8	-84.6	256
14837	**Madison**	WI	43.1	-89.3	262
14839	Milwaukee	WI	43.0	-87.9	211
14840	Muskegon	MI	43.2	-86.3	191
14842	Peoria	IL	40.7	-89.7	199
14847	Sault Ste. Marie	MI	46.5	-84.4	221
14848	South Bend	IN	41.7	-86.3	236
14850	*Traverse City*	MI	44.7	-85.6	192
14852	Youngstown	OH	41.3	-80.7	361

WBAN	Station	State	Lat.	Long.	Elev. (m)
14860	Erie	PA	42.1	-80.2	225
14891	Mansfield	OH	40.8	-82.5	395
14895	Akron	OH	40.9	-81.4	377
14898	Green Bay	WI	44.5	-88.1	214
14913	Duluth	MN	46.8	-92.2	432
14914	Fargo	ND	46.9	-96.8	274
14918	International Falls	MN	48.6	-93.4	361
14920	La Crosse	WI	43.9	-91.3	205
14922	Minneapolis	MN	44.9	-93.2	255
14923	Moline	IL	41.5	-90.5	181
14925	Rochester	MN	43.9	-92.5	402
14926	Saint Cloud	MN	45.6	-94.1	313
14933	Des Moines	IA	41.5	-93.7	294
14935	Grand Island	NE	41.0	-98.3	566
14936	Huron	SD	44.4	-98.2	393
14940	*Mason City*	IA	43.2	-93.3	373
14941	Norfolk	NE	42.0	-97.4	471
14943	Sioux City	IA	42.4	-96.4	336
14944	Sioux Falls	SD	43.6	-96.7	435
14991	*Eau Claire*	WI	44.9	-91.5	273
21504	Hilo	HI	19.7	-155.1	11
22516	Kahului	HI	20.9	-156.4	15
22521	**Honolulu**	HI	21.3	-157.9	5
22536	Lihue	HI	22.0	-159.4	45
23023	**Midland/ Odessa**	TX	31.9	-102.2	871
23034	San Angelo	TX	31.4	-100.5	582
23042	Lubbock	TX	33.7	-101.8	988
23044	**El Paso**	TX	31.8	-106.4	1194
23047	Amarillo	TX	35.2	-101.7	1098
23050	**Albuquerque**	NM	35.1	-106.6	1619
23061	**Alamosa**	CO	37.5	-105.9	2297
23065	Goodland	KS	39.4	-101.7	1124
23066	**Grand Junction**	CO	39.1	-108.5	1475
23129	Long Beach	CA	33.8	-118.2	17
23153	*Tonopah*	NV	38.1	-117.1	1653
23154	**Ely**	NV	39.3	-114.9	1906
23155	Bakersfield	CA	35.4	-119.1	150
23160	**Tucson**	AZ	32.1	-110.9	779
23161	**Daggett**	CA	34.9	-116.8	588
23169	**Las Vegas**	NV	36.1	-115.2	664
23174	**Los Angeles**	CA	33.9	-118.4	32
23183	**Phoenix**	AZ	33.4	-112.0	339
23184	*Prescott*	AZ	34.7	-112.4	1531
23185	Reno	NV	39.5	-119.8	1341
23188	**San Diego**	CA	32.7	-117.2	9
23232	Sacramento	CA	38.5	-121.5	8
23234	San Francisco	CA	37.6	-122.4	5
23273	**Santa Maria**	CA	34.9	-120.5	72
24011	**Bismarck**	ND	46.8	-100.8	502
24013	Minot	ND	48.3	-101.3	522
24018	Cheyenne	WY	41.2	-104.8	1872
24021	**Lander**	WY	42.8	-108.7	1696
24023	North Platte	NE	41.1	-100.7	849
24025	*Pierre*	SD	44.4	-100.3	526
24027	Rock Springs	WY	41.6	-109.1	2056
24028	Scottsbluff	NE	41.9	-103.6	1206
24029	Sheridan	WY	44.8	-107.0	1209
24033	Billings	MT	45.8	-108.5	1088

WBAN	Station	State	Lat.	Long.	Elev. (m)
24089	Casper	WY	42.9	-106.5	1612
24090	Rapid City	SD	44.1	-103.1	966
24121	Elko	NV	40.8	-115.8	1547
24127	**Salt Lake City**	UT	40.8	-112.0	1288
24128	Winnemucca	NV	40.9	-117.8	1323
24131	**Boise**	ID	43.6	-116.2	874
24143	**Great Falls**	MT	47.5	-111.4	1116
24144	Helena	MT	46.6	-112.0	1188
24146	Kalispell	MT	48.3	-114.3	904
24153	Missoula	MT	46.9	-114.1	972
24155	Pendleton	OR	45.7	-118.9	456
24156	Pocatello	ID	42.9	-112.6	1365
24157	Spokane	WA	47.6	-117.5	721
24221	**Eugene**	OR	44.1	-123.2	109
24225	**Medford**	OR	42.4	-122.9	396
24227	Olympia	WA	47.0	-122.9	61
24229	**Portland**	OR	45.6	-122.6	12
24230	**Redmond/ Bend**	OR	44.3	-121.2	940
24232	Salem	OR	44.9	-123.0	61
24233	**Seattle/ Tacoma**	WA	47.5	-122.3	122
24243	Yakima	WA	46.6	-120.5	325
24283	*Arcata*	CA	41.0	-124.1	69
24284	North Bend	OR	43.4	-124.3	5
25308	Annette	AK	55.0	-131.6	34
25339	Yakutat	AK	59.5	-139.7	9
25501	Kodiak	AK	57.8	-152.3	34
25503	King Salmon	AK	58.7	-156.7	15
25624	Cold Bay	AK	55.2	-162.7	29
25713	St Paul Is.	AK	57.2	-170.2	7
26411	**Fairbanks**	AK	64.8	-147.9	138
26415	Big Delta	AK	64.0	-145.7	388
26425	Gulkana	AK	62.2	-145.5	481
26451	Anchorage	AK	61.2	-150.0	35
26510	Mcgrath	AK	63.0	-155.6	103
26528	Talkeetna	AK	62.3	-150.1	105
26533	*Bettles*	AK	66.9	-151.5	205
26615	*Bethel*	AK	60.8	-161.8	46
26616	*Kotzebue*	AK	66.9	-162.6	5
26617	Nome	AK	64.5	-165.4	7
27502	*Barrow*	AK	71.3	-156.8	4
41415	**Guam**	PI	13.6	-144.8	110
93037	Colorado Springs	CO	38.8	-104.7	1881
93058	Pueblo	CO	38.3	-104.5	1439
93129	*Cedar City*	UT	37.7	-113.1	1712
93193	**Fresno**	CA	36.8	-119.7	100
93721	Baltimore	MD	39.2	-76.7	47
93729	**Cape Hatteras**	NC	35.3	-75.6	2
93730	Atlantic City	NJ	39.5	-74.6	20
93738	***Sterling (Washington-Dulles Airpt.)***	VA	39.0	-77.5	82
93805	**Tallahassee/ Apalachicola**	FL	30.4	-84.4	21
93814	Covington	KY	39.1	-84.7	271
93815	Dayton	OH	39.9	-84.2	306
93817	Evansville	IN	38.1	-87.5	118
93819	**Indianapolis**	IN	39.7	-86.3	246
93820	Lexington	KY	38.0	-84.6	301

WBAN	Station	State	Lat.	Long.	Elev. (m)
93821	Louisville	KY	38.2	-85.7	149
93822	Springfield	IL	39.8	-89.7	187
93842	Columbus	GA	32.5	-85.0	136
93987	*Lufkin*	TX	31.2	-94.8	96
94008	Glasgow	MT	48.2	-106.6	700
94018/ 23062	**Boulder/ Denver**	CO	39.8	-104.9	1610
94185	**Burns**	OR	43.6	-119.1	1271
94224	Astoria	OR	46.2	-123.9	7
94240	Quillayute	WA	48.0	-124.6	55
94702	Bridgeport	CT	41.2	-73.1	2
94725	*Massena*	NY	44.9	-74.9	63
94728/ 14732	**New York (LGA)**	NY	40.8	-73.9	11
94746	Worcester	MA	42.3	-71.9	301
94814	Houghton	MI	47.2	-88.5	329
94822	Rockford	IL	42.2	-89.1	221
94823	**Pittsburgh**	PA	40.5	-80.2	373
94830	Toledo	OH	41.6	-83.8	211
94846	Chicago	IL	41.8	-87.8	190
94847	Detroit	MI	42.4	-83.0	191
94849	Alpena	MI	45.1	-83.6	210
94860	Grand Rapids	MI	42.9	-85.5	245
94910	Waterloo	IA	42.6	-92.4	265
94918/ 14942	**Omaha**	NE	41.3	-95.9	298

References

1. Urban DJ, Cook NJ. 1986. Ecological Risk Assessment. Hazard Evaluation Division Standard Evaluation Procedure EPA 540/9-85-001. US EPA Office of Pesticide Programs, Washington, DC, USA.
2. NRC. 1983. *Risk Assessment in the Federal Government: Managing the Process.* National Research Council, Committee on the Institutional Means for Assessment of Risks to Public Health. National Academy Press, Washington, DC, USA.
3. Carsel RF, Mulkey LA, Lorber MN, Baskin LB. 1985. The pesticide root zone model (PRZM): A procedure for evaluation pesticide leaching threats to groundwater. *Ecol Model* 30:49-69.
4. Burns LA. 2000. Exposure Analysis Modeling System (EXAMS): User Manual and System Documentation. EPA/600/R-00/081. U. S. Environmental Protection Agency, Athens, GA, USA.
5. CENR. 1999. Ecological risk assessment under FIFRA. In *Ecological Risk Assessment in the Federal Government* CENR/5-99/001. National Science and Technology Council, Committee on Environment and Natural Resources (CENR), Washington, DC, USA, pp 3.1-3.11.
6. EPA. 1998. Guidelines for Ecological Risk Assessment. *Fed Regist* 63:26846-26924.
7. EPA. 1993. A Review of Ecological Assessment Case Studies from a Risk Assessment Perspective. Risk Assessment Forum EPA/630/R-92/005. U.S. Environmental Protection Agency, Washington, DC, USA.
8. SETAC. 1994. Final Report: Aquatic Risk Assessment and Mitigation Dialogue Group. Society of Environmental Toxicology and Chemistry, Pensacola, FL, USA.
9. NRC. 1994. *Science and Judgment in Risk Assessment.* National Research Council, Board on Environmental Studies and Toxicology. National Academy Press, Washington, DC, USA.
10. Hansen F. 1997. Policy for Use of Probabilistic Analysis in Risk Assessment at the U.S. Environmental Protection Agency. In *Memorandum dated May 15, 1997: Use of Probabilistic Techniques (Including Monte Carlo Analysis) in Risk Assessment.* U. S. Environmental Protection Agency, Washington, DC, USA.
11. Firestone M, Fenner-Crisp P, Barry T, Bennett D, Chang S, Callahan M, Burke A, Michaud J, Olsen M, Cirone P, Barnes D, Wood WP, Knott SM. 1997. Guiding Principles for Monte Carlo Analysis. EPA/630/R-97/001. U.S. Environmental Protection Agency, Washington, DC, USA.
12. Gallagher K, Touart L, Lin J, Barry T. 2001. A Probabilistic Model and Process to Assess Risks to Aquatic Organisms. Report prepared for May 13-16, 2001 FIFRA Scientific Advisory Panel Meeting; available at http://www.epa.gov/scipoly/sap/2001/march/aquatic.pdf U.S. Environmental Protection Agency, Washington, DC, USA.
13. SAP. 2001. Probabilistic Models and Methodologies: Advancing the Ecological Risk Assessment Process in the EPA Office of Pesticide Programs. Report of FIFRA Scientific Advisory Panel Meeting, March 13-16, 2001, held at the Sheraton Crystal City Hotel, Arlington, Virginia. SAP Report No. 2001-06, available at http://www.epa.gov/scipoly/sap/2001/march/march132001.pdf. US Environmental Protection Agency FIFRA Scientific Advisory Panel, Washington, DC, USA.
14. Oreskes N. 2000. Why believe a computer? Models, measures, and meaning in the natural world. In Schneiderman JS, ed, *The Earth Around Us: Maintaining a Livable Planet.* W. H. Freeman and Company, New York, NY, USA, pp 70-82.
15. Sunzenauer I. 1997. 1000 questions. E-mail dated 07/08/1997.
16. Oreskes N. 1998. Evaluation (not validation) of quantitative models. *Environ Health Perspect* 106 (Suppl. 6):1453-1460.
17. Maloszewski P, Zuber A. 1992. On the calibration and validation of mathematical models for the interpretation of tracer experiments in groundwater. *Advances in Water Resources* 15:47-62.
18. Mihram GA. 1972. Some practical aspects of the verification and validation of simulation models. *Operational Research Quarterly* 23:17-29.
19. Thomann RV. 1982. Verification of water quality models. *J Envir Eng Div, Proc ASCE* 108:923-940.
20. Oreskes N, Shrader-Frechette K, Belitz K. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* 263:641-646.
21. Gold HJ. 1977. *Mathematical Modeling of Biological Systems.* John Wiley & Sons, New York, NY, USA.
22. Popper KR. 1962. *Conjectures and Refutations: the Growth of Scientific Knowledge*. Basic Books, New York, NY, USA.
23. Popper KR. 1968. *The Logic of Scientific Discovery*. 3rd (revised) ed. Hutchinson, London, UK.

24. Kuhn TS. 1970. *The Structure of Scientific Revolutions*. 2nd ed, International Encyclopedia of Unified Science, Vol 2. University of Chicago Press, Chicago, IL, USA.
25. Cullen AC, Frey HC. 1999. *Probabilistic Techniques in Exposure Assessment: a Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. Plenum Press, New York, NY, USA.
26. IAEA. 1989. Evaluating the Reliability of Predictions Made Using Environmental Transfer Models. Safety Series 100. International Atomic Energy Agency, Vienna, Austria.
27. Kaplan S, Garrick BJ. 1981. On the quantitative definition of risk. *Risk Analysis* 1:11-27.
28. Ulam SM, von Neumann J. 1945. Random ergodic theorems. *Bulletin of the American Mathematical Society* 51:660.
29. Metropolis N, Ulam S. 1949. The Monte Carlo method. *Journal of the American Statistical Association* 44:335-341.
30. EPA. 1997. Exposure Factors Handbook. Volume I - General Factors. EPA/600/P-95/002Fa. U.S. Environmental Protection Agency, Washington, DC, USA.
31. Farrar D. 2001. Distributions for risk assessments: Some regulatory and statistical perspectives (in preparation). In *Pellston Workshop on the Application of Uncertainty Analysis to Ecological Risks of Pesticides*. SETAC Press, Pensacola, FL, USA.
32. Burns LA. 1983. Validation of exposure models: the role of conceptual verification, sensitivity analysis, and alternative hypotheses. In Bishop WE, Cardwell RD, Heidolph BB, eds, *Aquatic Toxicology and Hazard Assessment*, Vol ASTM STP 802. American Society for Testing and Materials, Philadelphia, PA, USA, pp 255-281.
33. Konikow LF, Bredehoeft JD. 1992. Ground-water models cannot be validated. *Advances in Water Resources* 15:75-83.
34. Anderson MP, Woessner WW. 1992. The role of the postaudit in model validation. *Advances in Water Resources* 15:167-173.
35. Mossman DJ, Schnoor JL. 1989. Post-audit study of dieldrin bioconcentration model. *Journal of Environmental Engineering* 115:675-679.
36. Beck MB. 1987. Water quality modeling: A review of the analysis of uncertainty. *Water Resour Res* 23:1393-1442.
37. Sargent RG. 1982. Verification and validation of simulation models. In Cellier FE, ed, *Progress in Modelling and Simulation*. Academic Press, London, UK, pp 159-169.
38. Rastetter EB. 1996. Validating models of ecosystem response to global change. *BioScience* 46:190-198.
39. Aldenberg T, Janse JH, Kramer PRG. 1995. Fitting the dynamic model PCLake to a multi-lake survey through Bayesian statistics. *Ecol Model* 78:83-99.
40. Balci O, Sargent RG. 1981. A methodology for cost-risk analysis in the statistical validation of simulation models. *Communications of the ACM* 24:190-197.
41. Beck MB, Ravetz JR, Mulkey LA, Barnwell TO. 1997. On the problem of model validation for predictive exposure assessments. *Stochastic Hydrology and Hydraulics* 11:229-254.
42. Flavelle P. 1992. A quantitative measure of model validation and its potential use for regulatory purposes. *Advances in Water Resources* 15:5-13.
43. Luis SM, McLaughlin D. 1992. A stochastic approach to model validation. *Advances in Water Resources* 15:15-32.
44. Lynch DR, Davies AM, eds. 1995. *Quantitative Skill Assessment for Coastal Ocean Models*, Coastal and Estuarine Studies Vol 47. American Geophysical Union, Washington, DC, USA.
45. Marcus AH, Elias RW. 1998. Some useful statistical methods for model validation. *Environ Health Perspect* 106 (Suppl. 6):1541-1550.
46. Mayer DG, Butler DG. 1993. Statistical validation. *Ecol Model* 68:21-32.
47. Ören TI. 1981. Concepts and criteria to assess acceptability of simulation studies: A frame of reference. *Communications of the ACM* 24:180-189.
48. Parrish RS, Smith CN. 1990. A method for testing whether model predictions fall within a prescribed factor of true values, with an application to pesticide leaching. *Ecol Model* 51:59-72.
49. Priesendorfer RW, Barnett TP. 1983. Numerical model-reality intercomparison tests using small-sample statistics. *Journal of the Atmospheric Sciences* 40.
50. Reckhow KH, Chapra SC. 1983. Confirmation of water quality models. *Ecol Model* 20:113-133.
51. Reckhow KH, Clements JT, Dodd RC. 1990. Statistical evaluation of mechanistic water-quality models. *Journal of Environmental Engineering* 116:250-268.
52. Shaeffer DL. 1980. A model evaluation methodology applicable to environmental assessment models. *Ecol Model* 8:275-295.
53. Venkatram A. 1982. A framework for evaluating air quality models. *Boundary-Layer Meteorol* 24:371-385.
54. Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR, O'Donnell J, Rowe CM. 1985. Statistics for the evaluation and comparison of models. *J Geophys Res* 90:8995-9005.
55. Mitchell PL. 1997. Misuse of regression for empirical validation models. *Agric Syst* 54:313-326.
56. Lassiter RR, Parrish RS, Burns LA. 1986. Decomposition by planktonic and attached microorganisms improves chemical fate models. *Environ Toxicol Chem* 5:29-39.
57. Shrader-Frechette KS. 1994. Science, environmental risk assessment, and the frame problem. *BioScience* 44:548-551.
58. Freund JE. 1962. *Mathematical Statistics*. Prentice-Hall, Englewood Cliffs, NJ, USA.
59. Diamond WJ. 1981. *Practical Experiment Designs for Engineers and Scientists*. Lifetime Learning Publications, Belmont, CA, USA.
60. USEPA. 1982. Testing for Field Applicability of Chemical Exposure Models. Workshop on Field Applicability Testing, 15-18 March 1982. Exposure Modeling Committee Report. US Environmental Protection Agency, Athens, GA, USA.

61. Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52:591-611.
62. Bratley P, Fox BL, Schrage LE. 1987. *A Guide to Simulation*. 2nd ed. Springer-Verlag, New York, NY, USA.
63. Zepp RG, Cline DM. 1977. Rates of direct photolysis in aquatic environment. *Environ Sci Technol* 11:359-366.
64. Iqbal M. 1983. *An Introduction to Solar Radiation*. Academic Press, New York, NY, USA.
65. Green AES, Schippnick PF. 1982. UV-B reaching the surface. In Calkins J, ed, *The Role of Solar Ultraviolet Radiation in Marine Ecosystems*. Plenum Press, New York, NY, USA, pp 5-27.
66. Teske ME, Bird SL, Esterly DM, Curbishley TB, Ray SL, Perry SG. 2001. AgDRIFT®: A model for estimating near-field spray drift. *Environ Toxicol Chem* (Accepted).
67. Bird SL, Perry SG, Teske ME. 2001. Evaluation of the AgDRIFT® aerial spray drift model. *Environ Toxicol Chem* (Accepted).
68. Jones RL, Russell MH, eds. 2000. *FIFRA Environmental Model Validation Task Force Final Report (Penultimate Draft)*.
69. Pollard JE, Hern SC. 1985. A field test of the Exams model in the Monongahela River. *Environ Toxicol Chem* 4:362-369.
70. Schramm K-W, Hirsch M, Twele R, Hutzinger O. 1988. Measured and modeled fate of Disperse Yellow 42 in an outdoor pond. *Chemosphere* 17:587-595.
71. Kolset K, Aschjem BF, Christopherson N, Heiberg A, Vigerust B. 1988. Evaluation of some chemical fate and transport models. A case study on the pollution of the Norrsundet Bay (Sweden). In Angeletti G, Bjørseth A, eds, *Organic Micropollutants in the Aquatic Environment (Proceedings of the Fifth European Symposium, held in Rome, Italy October 20-22, 1987)*. Kluwer Academic Publishers, Dordrecht, pp 372-386.
72. Kolset K, Heiberg A. 1988. Evaluation of the 'fugacity' (FEQUM) and the 'EXAMS' chemical fate and transport models: A case study on the pollution of the Norrsundet Bay (Sweden). *Water Sci Technol* 20:1-12.
73. Armbrust KL, Okamoto Y, Grochulska J, Barefoot AC. 1999. Predicting the dissipation of bensulfuron methyl and azimsulfuron in rice paddies using the computer model EXAMS2. *J Pest Sci* 24:357-363.
74. Tynan P, Watts CD, Sowray A, Hammond I. 1991. Field measurement and modelling for styrene, xylenes, dichlorobenzenes and 4-phenyldodecane. In *Proceedings of the 6th European Symposium on Aquatic Environment, 1990*, pp 20-37.
75. Games LM. 1982. Field validation of Exposure Analysis Modeling System (Exams) in a flowing stream. In Dickson KL, Maki AW, Cairns J, Jr., eds, *Modeling the Fate of Chemicals in the Aquatic Environment*. Ann Arbor Science Publ., Ann Arbor, Michigan, pp 325-346.
76. Sanders PF, Seiber JN. 1984. Organophosphorus pesticide volatilization: Model soil pits and evaporation ponds. In Kreuger RF, Seiber JN, eds, *Treatment and Disposal of Pesticide Wastes*. ACS Symposium Series, Vol 259. American Chemical Society, Washington, D.C., pp 279-295.
77. Barber MC, Suárez LA, Lassiter RR. 1988. Modeling bioconcentration of non-polar organic pollutants by fish. *Environ Toxicol Chem* 7:545-558.
78. Barber MC. 2002. A comparison of models for predicting chemical bioconcentration in fish. *Can J Fish Aquat Sci* (In preparation).
79. Barber MC, Suárez LA, Lassiter RR. 1991. Modelling bioaccumulation of organic pollutants in fish with an application to PCBs in Lake Ontario salmonids. *Can J Fish Aquat Sci* 48:318-337.
80. Barber MC. 2001. Bioaccumulation and Aquatic System Simulator (BASS). User's Manual Beta Test Version 2.1. EPA/600/R-01/035. U.S. Environmental Protection Agency, Office of Research and Development, Athens, GA, USA.
81. SCS. 1981. Land Resource Regions and Major Land Resource Areas of the United States. Agriculture Handbook 296. United States Department of Agriculture Soil Conservation Service, Washington, DC, USA.
82. Johnson GL, Hanson CL, Hardegree SP, Ballard EB. 1996. Stochastic weather simulation: Overview and analysis of two commonly used models. *Journal of Applied Meteorology* 35:1878-1896.
83. Hanson CL, Cumming KA, Woolhiser DA, Richardson CW. 1993. Program for Daily Weather Simulation. Water Resources Investigations Rep. 93-4018. U.S. Geological Survey, Denver, CO, USA.
84. Hanson CL, Cumming KA, Woolhiser DA, Richardson CW. 1994. Microcomputer Program for Daily Weather Simulation in the Contiguous United States. Agricultural Research Service ARS-114. U.S. Department of Agriculture, Boise, ID, USA.
85. Nicks AD, Gander GA. 1994. CLIGEN: A weather generator for climate inputs to water resource and other models. In *Proceedings of the Fifth International Conference on Computers in Agriculture*. American Society of Agricultural Engineers, Orlando, FL, USA.
86. Semenov MA, Brooks RJ, Barrow EM, Richardson CW. 1998. Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. *Clim Res* 10:95-107.
87. Richardson CW, Wright DA. 1984. WGEN: A Model for Generating Daily Weather Variables. Agricultural Research Service ARS-8. U.S. Department of Agriculture, Washington, DC, USA.
88. Semenov MA, Barrow EM. 1997. Use of a stochastic weather generator in the development of climate change scenarios. *Climate Change* 35:397-414.
89. McPeters RD, Bhartia PK, Krueger AJ, Herman JR, Schlesinger BM, Wellemeyer CG, Seftor CJ, Jaross G, Taylor SL, Swissler T, Torres O, Labow G, Byerly W, Cebula RP. 1996. Nimbus-7 Total Ozone Mapping Spectrometer (TOMS) Data Products User's Guide. NASA Reference Publication, National Aeronautics and Space

Administration, Scientific and Technical Information Branch, Lanham, MD, USA.
90. Aspelin AL. 1994. Pesticides Industry Sales and Usage: 1992 and 1993 Market Estimates. EPA 733-K-94-001. US EPA Office of Prevention, Pesticides and Toxic Substances, Washington, DC, USA.
91. Aspelin AL. 1997. Pesticides Industry Sales and Usage: 1994 and 1995 Market Estimates. EPA 733-K-97-002. US EPA Office of Prevention, Pesticides and Toxic Substances, Washington, DC, USA.
92. USDA. 1994. Agricultural Resources and Environmental Indicators. Agricultural Handbook 705. U.S. Department of Agriculture, Economic Research Service, Natural Resources and Environment Division, Washington, DC, USA.
93. Gianessi L, Anderson JE. 1995. Pesticide Use in U.S. Crop Production: National Summary Report. National Center for Food and Agricultural Policy, Washington, DC, USA.
94. NASS. 1993. Agricultural Chemical Usage: 1992 Field Crops Summary. Ag Ch 1(93). U.S. Department of Agriculture, National Agricultural Statistics Service and Economics Research Service, Washington, DC, USA.
95. NASS. 1993. Agricultural Chemical Usage: Vegetables–1992 Summary. Ag Ch 1(93). U.S. Department of Agriculture, National Agricultural Statistics Service and Economics Research Service, Washington, DC, USA.
96. NASS. 1994. Agricultural Chemical Usage: 1993 Fruits Summary. Ag Ch 1(94). U.S. Department of Agriculture, National Agricultural Statistics Service and Economics Research Service, Washington, DC, USA.